НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.942:519.216.3

До захисту допущено
В. о. завідувача кафедри ММСА
_____ О.Л.Тимощук
«___» _____ 2018 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 122 Комп'ютерні науки
на тему: «Методи data-mining в задачах прогнозування нелінійних нестаціонарних процесів»

Виконав:
студент ІІ курсу, групи КА-74мп
Саркісов Степан Юрійович _____

Керівник: професор кафедри ММСА,
д.т.н., професор,
Бідюк П.І. _____

Рецензент: професор кафедри інформаційної безпеки
НТУУ «КПІ ім. Ігоря Сікорського»,
д.т.н., професор,
Архипов О.Є. _____

Засвідчую, що у цій магістерській дисертації
немає запозичень з праць інших авторів
без відповідних посилань
Студент _____

Київ
2018

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність (спеціалізація) — 122 «Комп'ютерні науки» («Інтелектуальний аналіз даних в управлінні проектами»)

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри ММСА
_____ О.Л. Тимощук
«___» _____ 2018 р.

ЗАВДАННЯ
на магістерську дисертацію студенту Саркісову Степану Юрійовичу

1. Тема дисертації: «Методи data-mining в задачах прогнозування нелінійних нестаціонарних процесів», науковий керівник дисертації Бідюк Петро Іванович, професор, доктор технічних наук, затверджені наказом по університету від «07» листопада 2018 р. № 4121-с

2. Термін подання студентом дисертації: _____

3. Об'єкт дослідження: методи data-mining і прогнозування нелінійних нестаціонарних процесів.

4. Предмет дослідження: математичні моделі для формального опису нелінійних нестаціонарних процесів, критерії адекватності моделей і прогнозів на основі статистичних даних.

5. Перелік завдань, які потрібно розробити:
1) Огляд технічної літератури за темою роботи;
2) Дослідження актуальності обраної теми;
3) Вибір методів для моделювання і прогнозування;
4) Збір вхідних даних;
5) Виконання обчислювальних експериментів;
6) Аналіз результатів моделювання і прогнозування;
7) Проведення аналізу ринкових можливостей запуску стартап-проекту;
8) Підготовка ілюстративного матеріалу;
9) Оформлення пояснювальної записки.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:
1) Постановка завдання дослідження;
2) Методи інтелектуального аналізу даних;
3) Результати обчислень;
4) Наукова новизна результатів.

7. Орієнтовний перелік публікацій:

(1) XII Міжнародна науково-технічна конференція «Проблеми інформатизації», Київ, Державний університет телекомунікацій, грудень, 2018 р.

(2) Методи data-mining в задачах прогнозування нелінійних нестаціонарних процесів //Системні науки та кібернетика. – Стаття подана в редакцію журналу.

8. Дата видачі завдання: _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Отримання завдання на магістерську дисертацію	07.09.2018 – 09.09.2018	
2.	Огляд технічної літератури за темою	10.09.2018 – 30.09.2018	
3.	Дослідження актуальності обраної теми	01.10.2018 – 07.10.2018	
4.	Вибір методів для моделювання і прогнозування	08.10.2018 – 14.10.2018	
5.	Збір вхідних даних	15.10.2018 – 21.10.2018	
6.	Виконання обчислювальних експериментів	22.10.2018 – 28.10.2018	
7.	Аналіз результатів моделювання і прогнозування	29.10.2018 – 04.11.2018	
8.	Проведення аналізу ринкових можливостей запуску стартап-проекту	05.11.2018 – 11.11.2018	
9.	Підготовка ілюстративного матеріалу	12.11.2018 – 18.11.2018	
10.	Оформлення пояснювальної записки	19.11.2018 – 26.11.2018	

Студент С.Ю. Саркісов

Науковий керівник дисертації П.І. Бідюк

РЕФЕРАТ

Дипломна робота: 108 с., 27 рис., 27 табл., 27 джерел.

Тема дослідження: методи data-mining в задачах прогнозування нелінійних нестаціонарних процесів.

Об'єкт дослідження – статистичні дані стосовно розвитку досліджуваних процесів.

Предмет дослідження – методи статистичного аналізу масивів даних з метою побудови адекватних моделей досліджуваних процесів.

Мета роботи – підібрати підходящу модель для опису даних.

Метод дослідження – побудова математичних моделей вибраних процесів на основі масивів даних та оцінка статистичних критеріїв для перевірки адекватності побудованої моделі використовуючи технологію data mining.

Актуальність – створення системи, яка дозволить підібрати адекватну модель основану на критеріях.

Результати роботи – система, яка підбирає підходящі статистичні моделі.

Новизна роботи – запропоновано методику побудови регресійних моделей з використанням інформаційної технології data-mining; Запропонована модифікована процедура оцінювання структури моделі, яка відрізняється способом оцінювання умовної дисперсії; на основі статистичних даних побудовані нові моделі, які забезпечують можливість обчислення високоякісних оцінок прогнозів.

Шляхи подальшого розвитку предмету дослідження – вдосконалення обраних методів, розширення сфер використання.

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, КОМПЕТЕНЦІЯ, АВТОРЕГРЕСІЯ, ЧАСОВИЙ РЯД, ПРОГНОЗУВАННЯ.

ABSTRACT

Master’s thesis: 108 p., 27 fig., 27 tabl., 27 sources.

The topic of the research: methods of data-mining in the prediction of nonlinear nonstationary processes

Object of study – statistic’s data related to observed processes.

Purpose of the study – statisticals methods of analysing big data with purpose of making adequate models.

Purpose – chose the most suitable model for describing data.

Research method – making mathematical models chosing processes based on big data and evaluationg statisticals criterias for validating built models using data mining.

Urgency – creating system which allows to choose model based on criterias.

Results – system of choosing the most appropriate model.

The novelty of the work – proposed the method of formation regression models using data-mining technology; proposed modified procedure of structure assessment, based on assessing conditional variance; new models were created based on statistical data which give opportunity to calculate high quality forecast estimates.

The further development of the research subject – improvement of selected methods, expansion of use.

DECISION SUPPORT SYSTEMS, COMPETENCE, AUTOREGRESSION, TIME SERIES, FORECASTING.

Зміст

ПЕРЕЛІК ПРИЙНЯТИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

МНК – метод найменших квадратів;

САПП – середня абсолютна похибка в процентах;

СКП – сума квадратів похибок;

СПП – середня похибка в процентах;

ФЕП – фінансово-економічні процеси;

AIC – Akaike info criterion (інформаційний критерій Акайке);

ABT – Analytical Base Table;

BSC – Bias-Schwarz criterion (критерій Байєса-Шварца);

DW – Darbin-Wotson (статистика Дарбіна-Уотсона);

MAPE – mean absolute percent error (середня абсолютна похибка в процентах);

MAE – mean absolute error (середня абсолютна похибка);

R2 – коефіцієнт множинної детермінації;

RSME – root mean squared error (стандартне відхилення залишків, середньоквадратична помилка);

SSE – sum of squared errors (сума квадратів похибок);

ARCH – Autoregressive conditional heteroscedasticity (АРУГ - авторегресія з умовною гетероскедастичністю)

GARCH – Generalised autoregressive conditional heteroskedasticity (УАРУГ – узагальнена авторегресія з умовною гетероскедастичністю)

EGARCH – Exponential generalised autoregressive conditional heteroscedasticity (ЕУАРУГ – експоненційна узагальнена авторегресія з умовною гетероскедастичністю)

ВСТУП

Сучасний комп'ютерний термін Data Mining переводиться як «Вилучення інформації» або «Видобуток даних». Нерідко разом з Data Mining зустрічаються терміни Knowledge Discovery («Виявлення знань») і Data Warehouse («сховище даних»). Виникнення зазначених термінів, які є невід'ємною частиною Data Mining, пов'язане з новим витком у розвитку засобів і методів обробки та зберігання даних. Отже, мета Data Mining полягає у виявленні прихованих правил і закономірностей у великих (Дуже великих) обсягах даних.

Справа в тому, що людський розум сам по собі не пристосований для сприйняття величезних масивів різнорідної інформації. В середньому людина, за винятком деяких індивідуумів, не здатний уловлювати більше двох-трьох взаємозв'язків навіть у невеликих вибірках. Але і традиційна статистика, довгий час претендувала на роль основного інструмента аналізу даних, так само нерідко пасує при рішенні задач з реального життя. Вона оперує усередненими характеристиками вибірки, які часто є фіктивними величинами (середньої платоспроможністю клієнта, коли в залежності від функції ризику або функції втрат вам необхідно вміти прогнозувати спроможність і наміри клієнта; середньої інтенсивністю сигналу, тоді як вам цікаві характерні особливості та передумови піків сигналу і т. д.).

Тому методи математичної статистики виявляються корисними головним чином для перевірки заздалегідь сформульованих гіпотез, тоді як визначення гіпотези іноді буває досить складною і трудомісткою задачею. Сучасні технології Data Mining переробляють інформацію з метою автоматичного пошуку шаблонів (патернів), характерних для будь-яких фрагментів неоднорідних багатомірних даних. В відміну від оперативної аналітичної обробки даних (OLAP) в Data Mining тягар формулювання гіпотез і виявлення незвичайних (unexpected) шаблонів перекладено з людини на комп'ютер. Data Mining - це не один, а сукупність великого числа різних методів виявлення знань. Вибір методу часто

залежить від типу наявних даних і від того, яку інформацію ви намагаєтеся отримати. Ось, наприклад, деякі методи: асоціація (об'єднання), класифікація, кластеризація, аналіз часових рядів і прогнозування, нейронні мережі і т. д.

РОЗДІЛ 1 АКТУАЛЬНІСТЬ РОЗВ'ЯЗУВАННЯ ЗАДАЧ МОДЕЛЮВАННЯ І ПРОГНОЗУВАННЯ МЕТОДАМИ DATA-MINING

1.1 Актуальність застосування методології data-mining

Потенціал Data Mining дає "зелене світло" для розширення границь застосування технології. Щодо перспектив Data Mining можливі наступні напрямки розвитку:

- виділення типів предметних областей з відповідними їм евристиками, формалізація яких полегшить вирішення відповідних завдань Data Mining, що ставляться до цих областей;
- створення формальних мов і логічних засобів, за допомогою яких буде формалізовані міркування та автоматизація яких стане інструментом вирішення завдань Data Mining у конкретних предметних областях;
- створення методів Data Mining, здатних не тільки витягати з даних закономірності, але й формувати якісь теорії, що опираються на емпіричні дані;
- подолання істотного відставання можливостей інструментальних засобів Data Mining від теоретичних досягнень у цій області.

Якщо розглядати майбутнє Data Mining у короткостроковій перспективі, то очевидно, що розвиток цієї технології найбільш спрямовано до областей, пов'язаним з бізнесом.

У короткостроковій перспективі продукти Data Mining можуть стати такими ж звичайними й необхідними, як електронна пошта, і, наприклад, використання користувачами для пошуку найнижчих цін на певний товар або найбільш дешевих квитків.

У довгостроковій перспективі майбутнє Data Mining є дійсно захоплюючим – це може бути пошук інтелектуальними агентами як нових видів лікування різних захворювань, так і нового розуміння природи всесвіту.

Однак Data Mining таїть у собі й потенційну небезпеку – адже все більша кількість інформації стає доступнішою через всесвітню мережу, у тому числі й відомості приватного характеру, і усе більше знань можливо добути саме із неї:

Не дуже давно найбільший онлайновый магазин "Amazon" виявився в центрі скандалу із приводу отриманого їм патенту "Методи та системи допомоги користувачам при покупці товарів", що являє собою не що інше як черговий продукт Data Mining, призначений для збору персональних даних про відвідувачів магазина. Нова методика дозволяє прогнозувати майбутні запити на підставі фактів покупок, а також робити висновки про їхнє призначення. Ціль даної методики – те, про що говорилося вище – одержання як можна більшої кількості інформації про клієнтів, у тому числі й частки характеру (стать, вік, переваги і т.д.). Таким чином, збираються дані про приватне життя покупців магазина, а також членів їхніх родин, включаючи дітей. Останнє заборонено законодавством багатьох країн – збір інформації про неповнолітні можливий там тільки з дозволу батьків.

Дослідження відзначають, що існують як успішні рішення, що використають Data Mining, так і невдалий досвід застосування цієї технології. Області, де застосування технології Data Mining, швидше за все, будуть успішними, мають такі особливості:

- вимагають рішень, заснованих на знаннях;
- мають навколишнє середовище, що змінюється;
- мають доступні, достатні й значимі дані;
- забезпечують високі дивіденди від правильних рішень.

1.2 Методи, що використовуються в DATA-MINING

Всі методи Data Mining поділяються на дві великі групи за принципом роботи з вихідними навчальними даними. У цій класифікації верхній рівень

визначається на підставі того, зберігаються дані після Data Mining чи вони дистилюються для подальшого використання.

1. Безпосереднє використання даних, або збереження даних.

У цьому випадку вихідні дані зберігаються в явному деталізованому вигляді і безпосередньо використовуються на стадіях прогностичного моделювання та/або аналізу винятків. Проблема цієї групи методів - при їх використанні можуть виникнути складності аналізу надвеликих баз даних.

Методи цієї групи: кластерний аналіз, метод найближчого сусіда, метод k - найближчого сусіда, міркування за аналогією.

2. Виявлення і використання формалізованих закономірностей, або дистиляція шаблонів.

При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з вихідних даних і перетворюється в якісь формальні конструкції, вид яких залежить від використовуваного методу Data Mining. Цей процес виконується на стадії вільного пошуку, у першій же групі методів дана стадія в принципі відсутня. На стадіях прогностичного моделювання та аналізу винятків використовуються результати стадії вільного пошуку, вони значно компактніше самих баз даних. Нагадаємо, що конструкції цих моделей можуть бути трактовані аналітиком або не трактовані ("чорні ящики").

Методи цієї групи : логічні методи, методи візуалізації ; методи крос - табуляції; методи, засновані на рівняннях.

Логічні методи, або методи логічної індукції, включають:

- нечіткі запити і аналізи;
- символьні правила;
- дерева рішень;
- генетичні алгоритми.

Методи цієї групи є, мабуть, такими, що найкраще інтерпретуються - вони оформляють знайдені закономірності, в більшості випадків, у досить прозорому вигляді з точки зору користувача. Отримані правила можуть включати безперервні і дискретні змінні. Слід зауважити, що дерева рішень можуть бути

легко перетворені в набори символьних правил шляхом генерації одного правила по шляху від кореня дерева до його термінальної вершини. Дерева рішень і правила фактично є різними способами вирішення однієї задачі і відрізняються лише за своїми можливостями. Крім того, реалізація правил здійснюється більш повільними алгоритмами, ніж індукція дерев рішень.

Методи крос-табуляції: агенти, баесовські (довірчі) мережі, крос - таблична візуалізація. Останній метод не зовсім відповідає одній з властивостей Data Mining - самостійного пошуку закономірностей аналітичною системою. Однак, надання інформації у вигляді крос - таблиць забезпечує реалізацію основного завдання Data Mining - пошук шаблонів, тому цей метод можна також вважати одним з методів Data Mining.

Методи на основі рівнянь. Методи цієї групи висловлюють виявлені закономірності у вигляді математичних виразів - рівнянь. Отже, вони можуть працювати лише з чисельними змінними, і змінні інших типів повинні бути закодовані відповідним чином. Це дещо обмежує застосування методів даної групи, проте вони широко використовуються при вирішенні різних завдань, особливо завдань прогнозування.

Основні методи цієї групи: статистичні методи і нейронні мережі.

Статистичні методи найбільш часто застосовуються для вирішення задач прогнозування. Існує безліч методів статистичного аналізу даних, серед них, наприклад, кореляційно - регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз.

Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні та кібернетичні методи. Ця схема поділу заснована на різних підходах до навчання математичних моделей.

Слід зазначити, що існує два підходи віднесення статистичних методів до Data Mining. Перший з них протиставляє статистичні методи і Data Mining, його прихильники вважають класичні статистичні методи окремим напрямом аналізу даних. Відповідно до другого підходу, статистичні методи аналізу є частиною

математичного інструментарію Data Mining. Більшість авторитетних джерел дотримується другого підходу.

У цій класифікації розрізняють дві групи методів:

- статистичні методи, засновані на використанні усередненого накопиченого досвіду, який відображений в ретроспективних даних;
- кібернетичні методи, що включають безліч різнорідних математичних підходів.

Недолік такої класифікації: і статистичні, і кібернетичні алгоритми тим чи іншим чином спираються на зіставлення статистичного досвіду з результатами моніторингу поточної ситуації.

Перевагою такої класифікації є її зручність для інтерпретації - вона використовується при описі математичних засобів сучасного підходу до вилучення знань з масивів вихідних спостережень (оперативних і ретроспективних), тобто в задачах Data Mining.

Статистичні методи Data mining. Ці методи являють собою чотири взаємопов'язаних розділи:

- попередній аналіз природи статистичних даних (перевірка гіпотез стаціонарності, нормальності, незалежності, однорідності, оцінка виду функції розподілу, її параметрів тощо);
- виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз та ін);
- багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз та ін);
- динамічні моделі і прогноз на основі часових рядів.

Арсенал статистичних методів Data Mining класифікований на чотири групи методів:

1. Дескриптивний аналіз і опис вихідних даних.

2. Аналіз зв'язків (кореляційний та регресійний аналіз, факторний аналіз, дисперсійний аналіз).

3. Багатовимірний статистичний аналіз (компонентний аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції та ін.).

4. Аналіз часових рядів (динамічні моделі і прогнозування).

Кібернетичні методи Data Mining.

До цієї групи належать такі методи:

- Еволюційне програмування;
- Асоціативна пам'ять (пошук аналогів, прототипів);
- Нечітка логіка;
- Дерева рішень;
- Системи обробки експертних знань,
- Штучні нейронні мережі (розпізнавання, кластеризація, прогноз);
- Генетичні алгоритми (оптимізація).

Нейронні мережі (Neural Networks) - це клас моделей, що базуються на аналогії з роботою мозку людини і призначаються для вирішення різноманітних задач аналізу даних після проходження етапу навчання на даних.

Нейронні мережі - це моделі біологічних нейронних мереж мозку, в яких нейрони імітуються однотипними елементами (штучними нейронами).

Нейронна мережа може бути представлена направленим графом зі зваженими зв'язками, у якому штучні нейрони є вершинами, а синаптичні зв'язки - дугами.

Серед сфер застосування нейронних мереж - автоматизація процесів розпізнавання образів, прогнозування показників діяльності підприємства, медична діагностика, прогнозування, адаптивне управління, створення експертних систем, організація асоціативної пам'яті, оброблення аналогових і цифрових сигналів, синтез й ідентифікація електронних систем.

За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажу виробів, показники фінансового ринку, розпізнавати сигнали, конструювати самонавчальні системи.

Нейронна мережа є сукупністю нейронів, з яких складаються шари. У

кожному шарі нейрони пов'язані з нейронами попереднього і наступного шарів. Серед задач Data Mining, що вирішуються за допомогою нейронних мереж, розглядатимемо такі:

1. Класифікація (навчання з учителем). Приклади завдань класифікації: розпізнавання тексту, розпізнавання мови, ідентифікація особи.
2. Прогнозування. Для нейронної мережі задача прогнозування може бути поставленою так: знайти оптимальне наближення функції, заданої кінцевим набором вхідних значень.
3. Кластеризація (навчання без учителя). Прикладом задачі кластеризації може бути завдання стиснення інформації шляхом зменшення розмірності даних.
4. Генетичні алгоритми - різновид еволюційних обчислень. Засновником генетичних алгоритмів є Дж. Холланд. Суть їх розкривається у книзі "Адаптація у природних і штучних системах".

Генетичні алгоритми (ГА) - це алгоритми, що дають змогу знайти задовільне рішення для аналітично нерозв'язуваних проблем через послідовний підбір і комбінування параметрів з використанням механізмів, що нагадують біологічну еволюцію.

ГА належать до універсальних методів оптимізації, що дають змогу вирішувати задачі різних типів (комбінаторні, загальні задачі з обмеженнями і без обмежень) і різного ступеня складності. ГА характеризуються можливістю як однокри-теріального, так і багатокритеріального пошуку в інформаційному просторі. Інтеграція ГА і нейронних мереж допомагає вирішувати проблеми пошуку оптимальних значень ваг входів нейронів, а інтеграція ГА і нечіткої логіки дає можливість оп-тимізувати систему продукційних правил, які можуть бути використані для управління.

Різні методи Data Mining характеризуються певними властивостями. Серед основних властивостей і характеристик методів Data Mining можна назвати точність, масштабованість, здатність до інтерпретації, перевірки, трудомісткість, гнучкість, швидкість і популярність.

Масштабованість - властивість обчислювальної системи, що забезпечує розгорнення системних характеристик, наприклад, швидкості реакції, загальної продуктивності при додаванні до неї обчислювальних ресурсів.

Для досягнення успіху в інтелектуальному аналізі даних необхідно мати чітке уявлення про мету аналізу; зібрати реле-вантні дані; вибрати адекватні методи аналізу та перевірити передумови їх застосування; обрати програмно-технологічні та математичні засоби, що реалізують ці методи; виконати аналіз та прийняти рішення про використання результатів. Загальна схема використання методів Data Mining складається з таких етапів (рис 1.1):

Рисунок 1.1 - Етапи інтелектуального аналізу даних

На першому етапі виконується осмислення поставленої задачі і уточнення цілей, які мають досягатися методами Data Mining, тобто формується гіпотеза (рис. 1.1). Гіпотеза - частково обґрунтована закономірність знань, що слугує для зв'язку між різними емпіричними фактами або для пояснення факту чи групи фактів.

Важливо правильно сформулювати цілі і вибрати необхідні для їх досягнення методи, оскільки від цього залежить подальша ефективність усього процесу. Необхідно підібрати параметри, що якнайкраще описують об'єкт. Після вибору параметрів дані можуть бути представлені у вигляді таблиці. Після підготовки таблиці з описом параметрів потрібно оцінити значимість кожного з них. Можливо, частина з них буде відсіяна у результаті аналізу.

Є кілька методів збору необхідних для аналізу даних:

1) отримання цих даних з облікових систем;
2) отримання відомостей з непрямих даних;
3) використання відкритих джерел;
4) проведення власних маркетингових досліджень і заходів щодо збору даних;
5) збирання даних вручну.

Другий етап полягає у приведенні даних до форми, придатної для застосування методів Data Mining.

Третій етап - це застосування методів Data Mining, сценарії якого можуть бути різними і включати складну комбінацію різноманітних методів, особливо якщо методи дозволяють проаналізувати дані з різних позицій.

Наступний етап - перевірка побудованих моделей. Дуже простий і часто використовуваний спосіб полягає у тому, що всі наявні дані, які необхідно аналізувати, поділяються на дві групи різної розмірності. На більшій групі, застосовуючи методи Data Mining, одержують моделі, а на меншій - перевіряють їх. За різницею в точності між тестовою і навчальною групами можна стверджувати про адекватність побудованої моделі.

Останній етап - інтерпретація одержаних моделей експертом у цілях їх використання для прийняття рішень, додавання нових правил і залежностей у бази знань. Цей етап часто має на увазі використання методів, що знаходяться на стику технології Data Mining і технології експертних систем.

Методи Data Mining також можна класифікувати за задачами Data Mining. Відповідно до такої класифікації виділяємо дві групи. Перша з них - це підрозділ методів Data Mining на вирішальні завдання сегментації (тобто задачі класифікації та кластеризації) і завдання прогнозування.

У відповідності до другої класифікації по задачах методи Data Mining можуть бути спрямовані на отримання описових і прогнозуючих результатів.

Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з точки зору аналітика.

До методів, спрямованих на отримання описових результатів, відносяться ітеративні методи кластерного аналізу, в тому числі: алгоритм k - середніх, k - медіани, ієрархічні методи кластерного аналізу, самоорганізуються карти Кохонена, методи крос - табличної візуалізації, різні методи візуалізації та інші.

Прогнозуючі методи використовують значення одних змінних для передбачення / прогнозування невідомих (пропущених) або майбутніх значень інших (цільових) змінних.

До методів, спрямованих на отримання прогнозуючих результатів, відносяться такі методи: нейронні мережі, дерева рішень, лінійна регресія, метод найближчого сусіда, метод опорних векторів та ін.

1.3 Завдання DATA-MINING

Існує кілька умовних класифікацій задач Data Mining. Ми будемо говорити про чотири базових типа завдань.

1. Класифікація - це встановлення залежності дискретної вихідної змінної від вхідних змінних.
2. Регресія - це встановлення залежності безперервної вихідної змінної від вхідних змінних.
3. Кластеризація - це угруповання об'єктів (спостережень, подій) на основі даних, що описують властивості об'єктів. Об'єкти усередині кластера повинні бути схожими один на одного і відрізнятися від інших, які увійшли в інші кластери.
4. Асоціація - виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X слідує подія Y. Такі правила називаються асоціативними.

Вперше ця задача була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її називають аналізом ринкової корзини (market basket analysis).

Якщо ж нас цікавить послідовність подій, що відбуваються, то можна говорити про послідовних шаблонах - встановленні закономірностей між пов'язаними в часі подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X через час t послідує подія Y.

Крім перерахованих завдань, часто виділяють:

- - аналіз відхилень (deviation detection),
- - аналіз зв'язків (Link alalysis),
- - відбір значимих ознак (Feature selection),

хоча ці завдання межують з очищенням і візуалізацією даних.

У загальному випадку неважливо, яким саме алгоритмом буде вирішуватися завдання, головне - мати метод рішення для кожного класу задач.

Рішення переважної більшості бізнес-завдань зводиться до процесу KDD. Раніше були описані базові блоки, з яких збирається практично будь-який бізнес-рішення.

Мета застосування моделей в методах Data Mining - виявлення нових властивостей і закономірностей досліджуваних об'єктів і процесів.Тому

інформаційний підхід тут дуже до речі: модель повинна самостійно виявити в даних властиві їм закономірності (в більшості випадків раніше невідомі і приховані) і набути властивостей, необхідні для відображення цих закономірностей. Комплекс методів, використовуваних для створення таких моделей, називається машинним навчанням, а самі моделі - учнями. В основі машинного навчання лежить навчальна вибірка. Вона може бути або отримана як сукупність спостережень за розвитком об'єкта або процесу в минулому, або (що зустрічається рідше) створена експертом або аналітиком на основі деяких гіпотез, аналогій, особистого досвіду і навіть інтуїції.

В даному випадку означає, що, можливо, для навчання моделі ми будемо використовувати не всі наявні дані, а деяку їх підмножину, яка найбільш повно відображає шукані властивості і закономірності. Дані з навчальної вибірки послідовно пред'являються моделі, в результаті чого модель набуває необхідні властивості.

Цей процес називається навчанням. Він є итеративной процедурою.

1.4 Критеріальна база технології data-mining

1.4.1 Поняття структури математичної моделі

Введемо поняття структури математичної моделі, яке будемо використовувати надалі. Поняття структури моделі містить у собі наступні елементи (параметри):

1. Порядок моделі, тобто порядок диференціального, різницевого чи іншого рівняння, що використовується для опису динаміки процеса чи об'єкта. Наприклад, стохастичне різницеве авторегресійне (АР) рівняння другого порядку має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k). \quad (1.1)$$

Тобто, порядок цього різницевого рівняння визначається числом затриманих у часі значень зінної, що використовуються у правій частині рівняння для опису змінної в лівій частині. Стохастичним воно називається тому, що у правій частині присутня випадкова змінна $\varepsilon(k)$, призначення якої ми пояснимо трохи нижче.

2. Вимірність моделі. Вона визначається числом рівнянь, що використовуються для математичного описання об'єкта чи процесу. Процес, котрий описують одним рівнянням, називають одновимірним чи скалярним. Процес, котрий описують двома і більше рівняннями, називають багатовимірним. Хоча більшість процесів у природі є багатовимірними, часто обмежуються обновимірними моделями, виходячи з їх простоти та зручності застосування.

3. Наявність нелінійностей та їх характер. Визначити наявність нелінійностей – не завжди проста задача. Так, для механічних і деяких інших систем наявність нелінійностей можна визначити шляхом вивчення законів, закономірностей і особливостей їхнього функціонування. Наприклад, відомо, що для механічних систем характерною є наявність нелінійностей типу «люфт», «тертя», гістерезис.

При побудові регресійних моделей частіше зустрічаються нелінійності відносно змінних і нелінійності відносно параметрів. Прикладом нелінійності відносно змінних може бути поліноміальна регресія виду:

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \varepsilon(k). \quad (1.2)$$

Коефіцієнти цього рівняння можна оцінювати звичайним методом найменших квадратів (МНК) при належній побудові матриці вимірювань (вона

буде розглянута нижче). Нелінійність відносно параметрів зумовлена наявністю в моделі добутків коефіцієнтів, наприклад, у виглядді

$$y(k) = a_0 + a_1 a_2 x(k) + a_2 \exp(-bx(k)) + \varepsilon(k). \qquad (1.3)$$

Коефіцієнти (параметри) такої моделі неможливо оцінити за допомогою звичайного МНК, тому для розв‘язання цієї задачі використовують нелінійний МНК, метод максимальної правдоподібності чи інші методи нелінійного оцінювання.

4. Час запізнення реакції на виході об'єкта відносно вхідного сигналу. Запізнювання по входу (лаг) досить легко враховується як у неперервних, так і в дискретних моделях. Для моделі з дискретними змінними у вигляді різницевого рівняння

$$y(k) = a_0 + a_1 y(k-1) + a_2 x(k-d) + \varepsilon(k). \qquad (1.4)$$

час запізнення d представляє ціле число, що дорівнює числу періодів дискретизації вимірювань, на яке запізнюється вихідний сигнал щодо вхідного. Тривалість періоду дискретизації вимірювань залежить від динаміки конкретного процесу і може змінюватися в межах від декількох мікросекунд у фізико-технічних системах до одного року в макроекономіці.

Розглянемо модель неперервного процесу у вигляді передаточної функції із запізненням:

$$W(p) = \frac{K e^{-p\tau}}{1 + Tp}, \qquad (1.5)$$

де $K -$ статичний коефіцієнт передачі об’єкта;

$p -$ змінна Лапласа; $T -$ постінй часу;

τ – час запізнення по входу.

Запізнення у дискретній формі d і запізнення в неперервній формі τ зв'язані між собою наступним чином:

$$\hat{d} = \text{int}\,(\tau / T_s), \tag{1.6}$$

де $\hat{d}$ – оцінка часу запізнення в дискретній формі ($\hat{d} = 0, 1, 2, ...$).

5. Тип збурень, що діють на процес, і спосіб їх врахування. Під збуреннями розуміють вхідні впливи процесу, котрі створюють, як правило, негативні умови для його протікання і не використовуються з тих чи інших причин як керуючі. Збурення поділяють на детерміновані і стохастичні, а враховуються вони в адитивній чи мультиплікативній формі. Вище ми привели різницеві рівняння, у яких збурення $\varepsilon(k)$ входить в адитивній формі. Приклад мультиплікативної форми:

$$h(k) = \upsilon(k)[\alpha_0 + \alpha_1 h(k-1)], \tag{1.7}$$

де $\upsilon(k)$ – мультиплікативне збурювання.

Частіше всього збурення описують розподілами випадкових величин (статистичні моделі), але в окремих випадках його можна виміряти і описати функціонально (математичні моделі збурень). Наприклад, можна виміряти температуру навколишнього середовища, яка впливає на протікання реакції у хімічному реакторі і побудувати відповідну функціональну залежність температури від часу.

Вибір структури моделі, що адекватна процесу, – задача не проста і вирішується, як правило, ітераційно. Спочатку структуру моделі оцінюють наближено на підставі дослідження закономірностей протікання процесу, аналізу кореляційних функцій, візуального аналізу даних. При цьому вибирають декілька

найбільш ймовірних структур (кандидатів). Потім обчислюють оцінки параметрів моделей-кандидатів і вибирають кращу з них, використовуючи відповідні статистичні характеристики якості моделей.

Якщо жодна з моделей-кандидатів не може вважатися адекватною для конкретного застосування, то необхідно досліджувати на інформативність експериментальні дані, які можуть бути недостатньо інформативними для оцінювання моделі. У такому випадку може знадобитися повторний чи додатковий збір експериментальних даних і корегування структури моделі.

1.4.2 Два основних методи побудови математичних моделей

Основними методами побудови математичних моделей є

- структурний;
- функціональний.

Структурний метод передбачає моделювання внутрішнього механізму взаємодії змінних, відображає їх фактичні взаємозв’язки.

Критерієм правильності структурної моделі є однаковий характер поведінки основних змінних реального процесу і моделі.

Розглянемо, наприклад, зростання інфляції внаслідок випуску додаткової грошової маси. Оскільки логіка цього процесу досить проста і існують експериментальні (статистичні) дані, які ілюструють зростання інфляції, то можна постулювати, що інфляція описується диференціальним або різницевим рівнянням першого/другого порядку (рис.1.2).

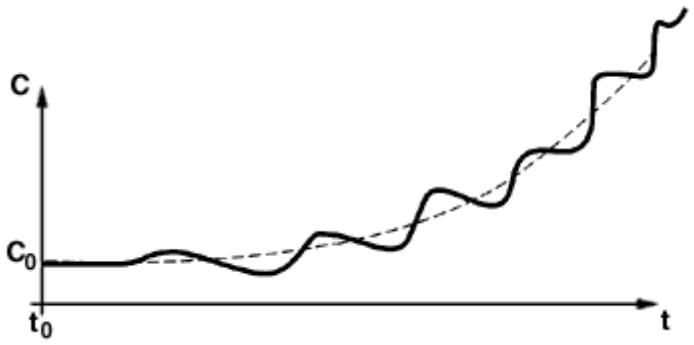

Рисунок 1.2 - Можлива крива зростання інфляції

Структурним підходом можна скористатись, наприклад, для побудови математичної моделі процесу трансформування власності (модель приватизації наведена в додатку) або макроекономіки в цілому. Для цього необхідно визначити вхідні керуючі змінні, збурення та вихідні змінні, а також визначити якого типу зв'язки існують між ними (рис.1.3).

Рисунок 1.3 - Спрощене зображення макроекономічного процесу

Серед керуючих змінних макроекономічного процесу можна виділити внутрішні та зовнішні інвестиційні потоки, потоки сировини, робочої сили, нові технології та структурні зміни в промисловості в цілому, а також в окремих галузях. Метою використання керуючих змінних є досягнення заданих рівнів макроекономічних показників – рівень виробництва валового внутрішнього продукту (ВВП), індекси інфляції, індекс людського розвитку, середня заробітна плата і т.і.

Як правило, в такі моделі вводять в явному вигляді збурення – випадкові змінні, які негативно впливають на протікання процесу. Так, у випадку створення моделі макроекономіки, збуреннями можуть бути

– помилкові рішення уряду;

– затримка платежів між підприємствами та державами;

– значні коливання цін на енергоносії;

– хронічна технологічна відсталість;

– швидкі зміни податкового законодавства;

– відплив капіталу за кордон;

– використання недостовірних статистичних даних.

Очевидно, що врахувати подібну інформацію в моделі надзвичайно складно, а тому випадкові змінні агрегують (об’єднують) і представляють в моделі однією чи двома випадковими змінними, які представляють всі збурення.

На основі знання логіки взаємодії змінних процесу та використання відомих макроекономічних законів (наприклад, врівноваженого розвитку процесів) будується система рівнянь, які описують розвиток окремої галузі або макроекономіки в цілому.

Функціональний підхід використовують для формального описання процесу, не проникаючи глибоко у фактичну структуру цього процесу і взаємодії його змінних.

Наприклад, для побудови моделі ціноутворення можна скористатись такими вхідними змінними: $I -$ об’єм імпорту; $D -$ об’єм грошової маси в обороті; $P -$ місячний об’єм виробництва. За вихідну змінною можна взяти індекс споживчих цін C, тобто,

$$C = f(I, D, P). \tag{1.8}$$

Для побудови цієї моделі необхідно мати чотири часових ряди, які необхідні для обчислення оцінок коефіцієнтів моделі.

Очевидно, що функціональний підхід є простішим від структурного і саме він найчастіше використовується на практиці. Гнучкість даного підходу дає можливість відносно швидко побудувати високоякісні моделі для прогнозування та синтезу систем керування.

1.4.3 Узагальнений алгоритм побудови моделі

Розглянемо узагальнений алгоритм побудови математичної моделі на основі експериментальних даних у вигляді наступних кроків. Він є узагальненим з точки зору застосування до систем чи процесів практично будь-якого типу.

1. Визначення мети побудови моделі, попереднє вивчення процесу (об’єкта).

На цьому етапі визначається мета побудови моделі, тобто чи буде модель використовуватись для поглибленого вивчення процесу, прогнозування його стану чи керування. Виконується аналіз функціонування процесу на основі літературних джерел та (можливо) експериментальних даних при їх наявності з метою встановлення числа входів і виходів, логіки взаємодії складових частин процесу, визначення можливих зовнішніх збурень та їх типу: детерміновані чи випадкові. По можливості необхідно встановити розподіл ймовірностей для випадкових збурень або функціональне описання для детермінованих збурень. Якщо існують моделі подібних процесів, їх також необхідно досконально вивчити та врахувати можливі недоліки.

2. Попередня оцінка структури моделі. На основі вивчення процесу, виконаного на першому етапі, необхідно встановити типи структур моделей-кандидатів. Їх може бути декілька в залежності від того наскільки невизначеною є інформація відносно процесу. Чим більшою є невизначеність, тим більше структур моделей необхідно досліджувати в процесі побудови адекватної моделі.

3. Планування експерименту та підготовка до його виконання. На цьому етапі виконуються наступні дії щодо планування виконання експерименту з метою отримання експериментальних даних:

– визначаються діапазони зміни вхідних та вихідних величин, збурень;

– встановлюється дискретність зміни вхідних величин, період дискретизації вимірів (якщо змінні неперервні), визначаються типи вимірювальних та реєструючих приладів;

– плануються режими роботи процесу, для яких необхідно зібрати експериментальні дані;

– якщо дані носять статистичний характер, то визначається періодичність їх збору та занесення в базу даних;

– задається тип, об’єм і якість продукції, що буде вироблена на протязі експерименту, а також визначаються необхідні об'єми сировини та енергії.

4. Виконання експерименту та формування бази даних. На цьому етапі реалізується розроблений на третьому етапі план експерименту і формуються часові ряди з вимірів (чи статистичних даних), які будуть використані для обчислювання оцінок параметрів математичних моделей.

5. Обчислення оцінок параметрів (коефіцієнтів) математичних моделей на основі експериментальних даних. При цьому оцінюють параметри для всіх моделей-кандидатів, вибраних на другому етапі. Для виконання цієї задачі необхідно:

– вибрати метод оцінювання параметрів моделі в залежності від її структури;

– зробити попередню обробку даних; в залежності від конкретної задачі це може бути масштабування, логарифмування, цифрова фільтрація, видалення недостовірних даних і т.п.;

– обчислити оцінки (векторів) параметрів моделей.

6. Визначити ступінь адекватності кожної моделі-кандидата процесу за допомогою статистичних критеріїв. Визначити кращу модель з множини кандидатів.

7. Якщо побудована модель відповідає висунутим вимогам (за точністю прогнозу чи якістю керування), то завершити процедуру; інакше перейти на 8-й крок.

8. Уточнити структуру моделі, зібрати, при необхідності, додаткові експериментальні дані і перейти на крок 5.

Хоча планування та виконання експерименту для соціально-економічних та фінансових систем є досить складною задачею, в окремих випадках це цілком можливо, особливо, якщо підприємство має наміри впровадити нові інформаційні технології обробки даних і методи прогнозування розвитку процесів на виробництві. Наприклад, цілком можливо спланувати та провести інвестиційний експеримент, експерименти з новими технологіями, новими типами продукції.

1.4.4 Вимоги до експериментальних даних, оцінок параметрів та моделі

1.4.4.1 Вимоги до експериментальних даних

1. Вимога неперервності та синхронності даних. Експериментальні дані повинні вимірюватись та реєструватись через однакові проміжки часу (період дискретизації вимірювань T_s). Цю вимогу необхідно виконувати для процесів будь-якого типу – технічних, економічних, екологічних і т.д. Порушення цієї вимоги призводить до зміни спектрального складу вимірювального сигналу, що недопустимо, оскільки при цьому змінюється інформативність сигналу. Крім того, вимірювання вхідних та вихідних сигналів необхідно робити синхронно, тобто в одні й ті ж моменти часу. В противному випадку вони будуть непридатні для побудови передаточних функцій, оскільки порушуються причинно-наслідкові зв’язки між входами та виходами. Як правило, в системах керування реального часу задача збору вимірювальних даних має найвищий пріоритет.

2. Вибірка даних повинна бути представницькою. Це означає, що вона повинна охоплювати досить довгий період часу, щоб включити в розгляд всі режими роботи, які передбачається описати моделлю. Розрізняють два основних

режими роботи процесів: перехідний та усталений. В перехідному режимі система керування переводить процес з деякого початкового стану в заданий. Перебування процесу в заданому (номінальному) стані на протязі деякого відносно довгого проміжку часу називають усталеним режимом роботи.

Прикладом широко відомого перехідного процесу є процес нагрівання до кипіння вмісту каструлі на кухонній плиті. На початку цього процесу ми задаємо режим максимальних витрат енергії щоб скоротити тривалість процесу нагрівання. Після досягнення режиму кипіння витрати енергії можна суттєво скоротити і процес переходить в усталений режим “повільного” кипіння. Подачу великої кількості енергії в початковий момент часу можна порівняти з подачею на вхід об’єкта сигналу у вигляді сходинки, а зміну температури вмісту каструлі можна вважати за перехідну характеристику цього процесу. Очевидно, що безліч прикладів такого типу можна знайти в промисловості. На сьогодні в нашому суспільстві спостерігається перехідний процес від соціалістичного ладу, який грунтувався на суспільній власності на засоби виробництва, до капіталістичного з приватною власністю.

3. Вибірка вимірювальних даних повинна бути інформативною. Частіше всього інформативність пов’язують з числом похідних, що їх містить вимірювальний сигнал. Чим більше число похідних можна отримати з вимірів, тим інформативнішим є сигнал. Наприклад, припустимо, що процес описується диференціальним рівнянням другого порядку:

$$a_2 \frac{d^2 y}{d t^2} + a_1 \frac{dy}{dt} + a_0 = b u(t), \tag{1.9}$$

де $y(t)$ – вихідний сигнал процесу;

$u(t)$ – вхідний сигнал;

$\theta^T = [a_0\ a_1\ a_2]$ – вектор коефіцієнтів рівняння, які необхідно оцінити за допомогою експериментальних даних.

Очевидно, що оцінки коефіцієнтів a_2 і a_1 можна обчислити тільки в тому випадку, якщо виміри $y(t)$ містять другу і першу похідну по часі.

Іноді інформативність визначають величиною дисперсії сигналу, тобто, чим більшою є дисперсія, тим вища інформативність сигналу. Так, константа має нульову дисперсію і, відповідно, мінімальну інформативність.

Вимога інформативності виконується в тому випадку, коли вхідний сигнал задовольняє умові достатнього збудження процесу. Основна ідея достатнього рівня збудження полягає в тому, щоб смуга частот вхідного сигналу перекривала амплітудно-частотну характеристику процесу. Тобто, вихідний сигнал $y(t)$ буде інформативним в тому випадку, коли достатньо інформативним буде вхідний сигнал $u(t)$. Умові інформативності (достатнього збудження) задовольняють такі основні вхідні сигнали: білий шум, псевдовипадковий двійковий сигнал та одиничний імпульс. Білий шум (гаусів процес) теоретично має нескінченний частотний спектр; достатньо широкі спектри мають і два інші сигнали.

З одного боку, для збудження процесу на його вхід необхідно подавати інформативний сигнал типу білого шуму, а з іншого – такий сигнал може бути недопустимим з точки зору фізики функціонування процесу (подача на вхід подібного сигналу може призвести до руйнування процесу чи до створення аварійної ситуації). Тому в таких системах часто використовують як збуджуючий сигнал завдання регулятору, якщо він має форму сходинки, тобто має фронт прямокутного імпульсу. В багатьох випадках можливе використання гармонічних сигналів, які сприймаються "легше" більшістю об'єктів ніж білий шум або одиничний імпульс. Так, при дослідженні механічних систем часто використовують одиничні імпульси, гармонічні збуджуючі сигнали та їх комбінації, а при дослідженні технологічних процесів до вхідного сигналу керування додають 10-15% білого шуму, який забезпечує достатній рівень „збудження” процесу.

1.4.4.2 Вимоги до оцінок параметрів моделі

Точність оцінок параметрів моделі залежить від якості вимірювальних даних, коректності попередньої обробки даних та від того, наскільки правильно вибрано метод оцінювання. Так, для оцінювання параметрів лінійних та псевдолінійних (нелінійних відносно змінних) моделей можна застосовувати звичайний МНК та його модифікації, а для оцінювання моделей, нелінійних відносно параметрів, необхідно застосовувати нелінійний МНК, метод максимальної правдоподібності та інші методи, розроблені для оцінювання параметрів нелінійних моделей.

Існують такі стандартні вимоги до оцінок параметрів математичних моделей:

1. Оцінки повинні бути незміщеними. Це означає, що оцінки параметрів не повинні містити систематичної похибки, яка збільшує або зменшує оцінки параметрів на всіх вибірках даних або на різних відрізках однієї вибірки. Формально незміщеність оцінок параметрів записують так:

$$E[\hat{\theta}]=\theta, \tag{1.10}$$

де $E-$ символ математичного сподівання;

$\hat{\theta}-$ вектор оцінок параметрів;

$\theta-$ істинне значення вектора параметрів.

2. Оцінки повинні бути консистентними, тобто оцінка $\hat{\theta}$ вектора параметрів повинна наближатись до свого істинного значення θ по мірі збільшення об’єму вибірки даних. Оскільки оцінка $\hat{\theta}$ – це випадкова величина, то

наближення до істинного значення можливе тільки в імовірнісному смислі. Консистентна оцінка повинна задовольняти наступному співвідношенню:

$$p\left(\left|\hat{\theta}_k - \theta\right| < \varepsilon\right) \to 1 \quad \text{при} \quad k \to \infty, \tag{1.11}$$

де $\varepsilon > 0$ – мале число;

$p -$ символ ймовірності;

$\hat{\theta}_k -$ оцінка вектора параметрів в момент k.

Відомо, що довжина перехідного процесу при оцінюванні моделі залежить від кількості параметрів, що оцінюються, та вимірності моделі, а тому об'єми вибірок даних в усіх випадках повинні бути, по можливості, більшими. Проблеми з необхідними об'ємами даних виникають, як правило, при моделюванні економічних та соціальних систем; при побудові моделей технічних систем такі проблеми досить рідкісні.

3. Оцінки повинні бути ефективними, а це означає, що із множини допустимих, незміщених та консистентних оцінок необхідно вибрати ті, що є найближчими до оцінюваних параметрів, тобто ті, що мають найменші відхилення від середнього значення.

Іншими словами це вимога мінімальності дисперсії оцінки, яка формально записується так:

$$Var(\hat{\theta}) \to \min. \tag{1.12}$$

Незміщені ефективні оцінки параметрів лінійної моделі можна отримати, наприклад, за допомогою методу найменших квадратів, якщо при оцінювання виконуються наступні умови:

– похибка моделі $e(k) = y(k) - \hat{y}(k)$ є центрованою величиною; де $y(k) -$ значення ряду, отримане експериментально (або статистичні дані); $\hat{y}(k) -$ оцінка змінної, отримана за допомогою побудованої моделі;

– похибка моделі – це некорельований процес, тобто, відсутня автокореляція похибок:

$$\operatorname{cov}[e(k)] = E[e(k)e(k-l)] = \begin{cases} \sigma_e^2, & k = l, \\ 0, & k \neq l. \end{cases} \tag{1.13}$$

– похибка моделі некорельована із залежною змінною $y(k)$.

Корельованість похибки означає, що вона містить інформацію процес. Тобто необхідно коригувати структуру моделі таким чином, щоб похибка стала некорельованою.

Постановка задачі і висновки до розділу

На даний момент data-mining є однієї з найпреспективніших методології нашого часу, тому що вона дає змогу оперувати з великою кількістю інформації і допомагає отримувати результати після її використання. Були розглянуті завдання data-mining, в яких сферах її можна використовувати та методи які дозволяють це зробити.

Ми детально розглянули схему ІАД та поетапні кроки які в результаті дають нам, ту модель, яка підходить для використання з нашими даними. Був проведений опис статистичних моделей як лінійних та і не лінійних, іх задач та критеріальної бази, яка на виході слугує нашим помічником у встановленні коректності побудови моделі.

Як моделі піддаються оцінці так і прогнози, але в даному випадку прогнози дуже чутливі від об’єму даних, на великих масивах ми можемо отримувати похибки та не завжди деякі критерії підходять для оцінки.

Тому при оцінці прогнозу/моделей доцільно використовувати декілька критеріїв та можливу іх варіації та комбінацію.

Постановка задачі:

1. Виконати огляд застосувань методики data-mining з використанням статистичних методів.

2. Вибрати типи математичних моделей.

3.Застосувати розроблену СППР до аналізу вибраних моделей.

3.1.Використати методи аналізу статистичних моделей. 3.2.Обчислити оцінки криетріїв до них.

4. Виконати порівняльний аналіз використаних моделей та критеріїв до них з метою вибору кращого для даної моделі, використовуючи автоматизований режим порівняння.

5. Виробити рекомендації стосовно можливостей розробленої системи до аналізу моделей.

РОЗДІЛ 2 СТРУКТУРИ МОДЕЛЕЙ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ

2.1 ЛІНІЙНІ ТА НЕЛІНІЙНІ ТРЕНДИ

2.1.1 ПОЛІНОМІАЛЬНІ, ЛІНІЙНІ, ГІПЕРБОЛІЧНІ МОДЕЛІ

Оцінка моделей, нелінійних по пояснюючим змінним, але лінійних по оцінюваним параметрами не представляє особливої складності: в цьому випадку зазвичай використовують заміну змінних для зведення моделі до лінійної і оцінки параметрів за допомогою звичайного МНК (застосованого до моделі з замінених змінними).

Так, у разі поліноміальною залежності ступеня k:

$$y = a_0 + a_1 x + a_2 x^2 + K + a_k x^k + \varepsilon \quad (2.1)$$

за допомогою заміни змінних:

$$z_1 = x, z_2 = x^2, K, z_k = x^k \quad (2.2)$$

Отримуємо лінійну модель множинної регресії з k пояснюючими змінними:

$$y = a_0 + a_1 z_1 + a_2 z_2 + K + a_k z_k + \varepsilon \quad (2.3)$$

Оцінки параметрів цієї лінійної моделі знаходять за допомогою звичайного МНК.

На практиці серед подібних поліноміальних регресій найбільш часто зустрічаються поліноми другого ступеня (квадратична або параболічна регресія):

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon \quad (2.4)$$

і третього ступеня (кубічна регресія):

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \varepsilon \quad (2.5)$$

Так, квадратична функція може відображати залежність між обсягом випуску і витратами (середніми або граничними) або між витратами на рекламу і прибутком. Зазвичай ця функція використовується тоді, коли всередині розглянутого інтервалу зміни фактора прямий або зворотний характер залежності змінюється на протилежний. Кубічна функція в макроекономіці відображає залежність загальних витрат від обсягу випуску. Поліноміальні функції добре підходять для моделювання ефекту масштабу, аналізу максимумів і мінімумів.

Модель виду:

$$y = \alpha + \frac{\beta}{x} + \varepsilon \quad (2.6)$$

називається зворотної (гіперболічної) моделлю.

Ця модель зводиться до лінійної з допомогою заміни:

$$z = \frac{1}{x} \quad (2.7)$$

Дана модель зазвичай застосовується в тих випадках, коли необмежене збільшення пояснюючої змінної x асимптотично наближає залежну змінну y до деякого межі. Обернені функції добре підходять для моделювання ефектів повного насичення і обмеженості. Залежно від знаків коефіцієнтом можна виділити наступні характерні випадки (рис 2.1):

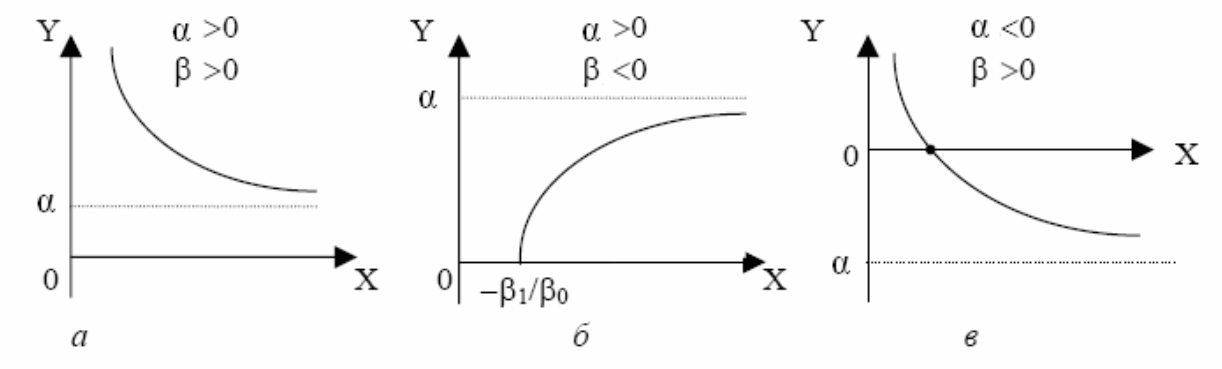

Рисунок 2.1 – залежність факторів

Залежність, зображена на рис. 2.1. а - може відображати залежності між обсягом випуску і середніми фіксованими витратами. Графік, зображений на рис. 2.1, б - залежність між доходом і попитом на блага (наприклад, на товари першої необхідності, які товари відносної розкоші) - так звані функції Торнквиста. Прикладом такої залежності такого виду можуть служити також криві Енгеля, що відображають взаємозв'язок між частки витрат на товари тривалого користування і загальних сум витрат (або доходів).

Важливим випадком графіка, зображеного на рис. 2.1, в - є крива Філліпса, що відображає залежність між відсотковим вимірюванням заробітної плати від рівня безробіття, вираженого у відсотках.

Модель виду:

$$y = \frac{1}{\alpha+\beta x+\varepsilon} \tag{2.8}$$

також є зворотною моделлю і може бути приведена до лінійної моделі: звертаючи обидві частини рівності, отримуємо лінійну формулу щодо змінної $1/y$:

$$1/y = \alpha + \beta x + \varepsilon \tag{2.9}$$

Можливі й інші моделі, нелінійні по пояснює пере- менним, які лінеарізуются заміною змінних.

Наприклад, лінійно логарифмічні (полулогарифмічні) залежності:

$$y = \alpha + \beta \ln x + \varepsilon \tag{2.10}$$

які приводяться до лінійної формі заміною:

$$z = \ln x \tag{2.11}$$

Подібні залежності також використовуються при моделюванні кривих Енгеля і характеризуються тим, що логарифм при пояснюючій змінній знижує вплив зростання цієї змінної (ступінь впливу x знижується з ростом x). Таким чином можна моделювати ефекти насичення на рівні швидкості росту: «зростання з порядку спадання швидкістю».

За допомогою заміни змінних можлива також оцінка залежностей з квадратними коренями, наприклад:

$$y = \alpha + \beta\sqrt{x} + \varepsilon \tag{2.12}$$

які приводяться до лінійної формі заміною:

$$z = \sqrt{x} \tag{2.13}$$

Неважко бачити, що довільна комбінація залежностей, наведених вище, може бути лінеаризована за допомогою відповідних замін змінних.

Кілька більш складним випадком є оцінка параметрів в разі нелінійності моделі за параметрами, так як безпосередній застосування МНК для їх оцінювання неможливо. В цьому випадку відповідним перетворенням (зазвичай пов'язаних з логарифмування по основі e) іноді вдається привести модель до лінійного вигляду.

Так, в разі статичної залежності:

$$y = \alpha x^{\beta} \varepsilon \tag{2.14}$$

Прологорифмував обидві частини, отримаємо:

$$\ln y = \ln \alpha + \beta \ln x + \ln \varepsilon \tag{2.15}$$

Отримана лінійна модель, в якій і залежна і пояснююча змінні задані в логарифмічному вигляді іноді називається подвійною логарифмічною моделлю.

2.1.2 Моделі процесів з детермінованим трендом

Тренд (поточне середнє) може бути зростаючим або спадаючим, а за характером зміни в часі може бути детермінованим або стохастичним.

Детермінований тренд описують вибраною функцією, наприклад, поліномом від часу, сплайном, експонентою, комбінацією тригонометричних функцій та інше. Часто використовують поліноми від часу вигляду:

$$y(k) = a_0 + a_1 \cdot k + a_2 \cdot k^2 + \ldots + a_m \cdot k^m + \varepsilon(k), \tag{2.16}$$

де k – дискретний час, який зв'язаний з неперервним реальним часом t через період реєстрації (дискретизації) даних: $t = kT_s$;

$\varepsilon(k)$ – випадкова змінна, оцінку якої можна знайти після оцінювання рівняння: $\hat{\varepsilon}(k) = e(k)$, де $e(k)$ – похибка моделі.

Очевидно, що після оцінювання моделі послідовність значень $\{e(k)\}$ буде містити всі коливання, що накладаються на тренд.

Окрім поліномів від часу, для описання тренду в моделюванні

використовують експоненціальні та гармонічні поліноми.

Випадкові тренди, тобто тренди, які не можна описати з необхідною точністю за допомогою детермінованих функцій, моделюють за допомогою випадкових процесів. В даній роботі цей підхід не розглядається.

Таким чином, описуючи тренд рівнянням (2.16), ми фактично видаляємо його з процесу і повна модель процесу буде складатись щонайменше з двох рівнянь: рівняння (1.1) для тренду і рівняння АР(p) або АРКС(p,q), яке описує коливання, що накладаються на тренд.

Тренд може бути видалений з процесу (даних) за допомогою різниць. Так, перші різниці видаляють тренд першого порядку (лінійний тренд), другі різниці видаляють квадратичний тренд і т.д. Наприклад, нехай $y(k) = a_0 + a_1 \cdot k$. Перші різниці цього процесу:

$$\Delta y(k) = y(k) - y(k-1) = a_0 + a_1 \cdot k - [a_0 + a_1 \cdot (k-1)] = a_1 \qquad (2.17)$$

приводять до видалення лінійного тренду. Очевидно, що після видалення тренду ми вже не зможемо його спрогнозувати. Докладно задача моделювання процесів з трендом буде розглянута в подальшому.

2.1.3 Тест на тренд

При перевірці на стаціонарність спочатку необхідно візуально дослідити часовий ряд. Нагадаємо, що слабка стаціонарність (яка частіше всього використовується на практиці) означає, що середнє значення, дисперсія та коваріація ряду не змінюються в часі. Досить часто вже попереднє візуальне дослідження дозволяє визначити присутність лінійного чи нелінійного тренду. Стаціонарний ряд має нульовий порядок інтегрованості, що формально записується так: $\{y(k)\}\ I(0)$.

Порядком інтегрованості є число, яке показує скільки разів необхідно застосувати до часового ряду оператор перших різниць, щоб перейти до стаціонарного ряду.

За визначенням часовий ряд має одиничний корінь або порядок інтеграції 1, тобто $\{y(k)\} I(1)$, якщо його перші різниці $\Delta y(k) = y(k) - y(k-1)$ утворюють стаціонарний ряд $\{\Delta y(k)\} I(0)$.

Часовий ряд має два одиничних корені або порядок інтеграції 2, якщо для досягнення стаціонарності необхідно обчислити його другі різниці:

$$\Delta^2 y(k) = \Delta y(k) - \Delta y(k-1) = y(k) - y(k-1) - [y(k-1) - y(k-2)]$$
$$= y(k) - 2y(k-2) + y(k-2), \qquad (2.18)$$

де $\{\Delta^2 y(k)\} I(0)$.

В загальному випадку часовий ряд може мати довільний порядок інтегрованості $\{y(k)\} \square I(\text{int})$. Для визначення існування нестаціонарності (існування одиничних коренів) запропоновано ряд тестів.

2.1.4 Перевірка присутності нестаціонарності (тест Дікі-Фуллера).

Після виконання візуального контролю необхідно застосувати формальні тести на стаціонарність, які дають можливість переконатись в її існуванні. Розглянемо порядок застосування тесту Дікі-Фуллера.

За допомогою цього критерію визначають яку величину має коефіцієнт a_1 в рівнянні:

$$y(k) = a_1 y(k-1) + \varepsilon(k), \tag{2.19}$$

тобто, $a = 1$ чи $a < 1$. Якщо $a = 1$, то дані містять одиничний корінь і степінь інтегрованості дорівнює буде $I(1)$. Якщо ж $0 < a_1 < 1$, то ряд стаціонарний, тобто має степінь інтегрованості $I(0)$. Для фінансово-економічних процесів значення $a_1 > 1$ не є характерним, тому що такі значення означають присутність в процесах різко зростаючих (спадаючих) ефектів. Виникнення таких процесів є малоймовірним, оскільки фінансово-економічне середовище є достатньо інерційним і не дозволяє змінним приймати нескінченно великі значення за короткі проміжки часу.

Нагадаємо, що застосування МНК для оцінювання коефіцієнтів моделі часового ряду передбачає, що залишки (похибки) $e(k)$ моделі мають постійну скінченну дисперсію. Присутність нестаціонарності приводить до порушення цього припущення. Наприклад, розглянемо рівняння:

$$\begin{aligned} y(k) &= y(k-1) + e(k) = [y(k-2) - e(k-1)] + e(k) = \ldots \\ &= y(0) + e(k) + e(k-1) + e(k-2) + \ldots + e(1). \end{aligned} \tag{2.20}$$

Оскільки залишки $e(k)$ незалежні і мають постійну дисперсію, то дисперсія $y(k)$ зростає до нескінченності при $k \to \infty$. В такому випадку для описання динаміки ряду можна скористатись рівнянням:

$$\Delta y(k) = b\, y(k-1) + e(k), \tag{2.21}$$

де $b = a_1 - 1$.

Якщо $b = 0$, то ряд містить одиничний корінь і має степінь інтегрованості $I(1)$, а ряд $\{\Delta y(k)\}$ може бути вже стаціонарним. Якщо ж $b < 0$, то $a < 1$ і стаціонарним буде сам ряд $\{y(k)\}$.

В рівнянні $y(k) = a_1 y(k-1) + \varepsilon(k)$ відсутнє середнє значення і описання тренда.

Якщо включити середнє, то воно приймає вигляд:

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k) \tag{2.22}$$

або

$$\Delta y(k) = a_0 + a_1 y(k-1) - y(k-1) + \varepsilon(k) = a_0 + by(k-1) + \varepsilon(k). \tag{2.23}$$

Із врахуванням тренда останнє рівняння приймає вигляд:

$$y(k) = a_0 + a_1 k + a_2 y(k-1) + \varepsilon(k), \tag{2.24}$$

де $k-$ дискретний час. Це рівняння можна записати для першої різниці

$$y(k) - y(k-1) = a_0 + a_1 k + by(k-1) - y(k-1) + \varepsilon(k) \tag{2.25}$$

або

$$\Delta y(k-1) = a_0 + a_1 k + by(k-1) + \varepsilon(k). \tag{2.26}$$

Для такої моделі було б некоректно використовувати $t-$статистику з метою визначення значимості коефіцієнта b, оскільки застосування регресії для оцінювання цього коефіцієнта передбачає, що $b < 0$ $(a_1 < 1)$. Тобто при $b \approx 0$ великий процент оцінок за $t-$статистикою не буде прийматися як значимий, тобто нульова гіпотеза щодо існування одиничного кореня буде часто відкидатись.

Крім того, одиничні корені робастні (зберігаються і можуть бути виявлені) при різних степенях гетероскедастичності, але можуть виникати проблеми з автокореляцією залишків моделі. В умовах присутності автокореляції залишків

задача тестування на стаціонарність розв'язується за допомогою розширеного тесту Дікі-Фуллера. При використанні цього методу значення залежної змінної вводяться в рівняння регресії з великими значеннями лагу, достатніми для того щоб уникнути автокореляції залишків. Це рівняння може мати наступний вигляд:

$$\Delta y(k) = a_0 + by(k-1) + c_1\,\Delta y(k-1) + c_2\,\Delta y(k-2) + \ldots + c_n\,y(k-n) + \varepsilon(k). \qquad (2.27)$$

Форма критерію значимості залежить від виду моделі, що тестується, тобто, чи включено в модель середнє значення і член, який описує тренд.

Нульова гіпотеза без середнього.

При тестуванні рівняння:

$$\Delta y(k) = b\,y(k-1) + c_1\,\Delta y(k-1) + c_2\,\Delta y(k-2) + \ldots + c_n\,y(k-n) + \varepsilon(k), \qquad (2.28)$$

тобто, середнє відсутнє, гіпотеза записується так:

$H_0 : b = 0$ - ряд нестаціонарний;

$H_1 : b < 0$ - ряд стаціонарний.

Нульова гіпотеза відкидається, якщо статистика b / SE_b має від'ємне значення, яке менше за критичне значення, взяте з таблиці Дікі-Фуллера. Критичні значення для рівнів значимості $\alpha = 1$ і $\alpha = 5$ дорівнюють –2,58 і –1,95, відповідно.

Якщо нульова гіпотеза приймається, то ряд $\{y(k)\}$ - це випадкове блукання без зсуву (константи в рівнянні).

В більш загальному вигляді цього критерію враховується розмір вибірки N, що досягається шляхом обчислення модифікованого критичного значення по формулі

$$\tau_\infty + \frac{\tau_1}{N} + \frac{\tau_2}{N^2},$$

де $\tau_\infty = -2{,}57$ $(\alpha = 1)$ або $\tau_\infty = -1{,}94$ $(\alpha = 5)$;

$\tau_1 = -1{,}96$ $(\alpha = 1)$ або $\tau_1 = -0{,}398$ $(\alpha = 5)$;

$\tau_2 = -10{,}04$ $(\alpha = 1)$ або $\tau_2 = 0$ $(\alpha = 5)$;

(значення τ табульовані Маккінноном, 1991).

Нульова гіпотеза з середнім значенням

Перевірка рівняння $\Delta y(k) = a_0 + b\, y(k-1) + e(k)$ із врахуванням можливої автокореляції залишків (як це було показано вище) базується на використанні того ж статистичного критерію, що і для рівняння без середнього, і тієї ж формули критичних значень, але при наступних значеннях τ:

$\tau_\infty = -3{,}43$ $(\alpha = 1)$ або $\tau_\infty = -2{,}86$ $(\alpha = 5)$;

$\tau_1 = -6{,}00$ $(\alpha = 1)$ або $\tau_1 = -2{,}74$ $(\alpha = 5)$;

$\tau_2 = -29{,}25$ $(\alpha = 1)$ або $\tau_2 = -8{,}36$ $(\alpha = 5)$.

Нульова гіпотеза при наявності середнього та тренду.

В даному випадку застосовується така ж процедура, що і вище, але при наступних значеннях τ:

$\tau_\infty = -3{,}96$ $(\alpha = 1)$ або $\tau_\infty = -3{,}41$ $(\alpha = 5)$;

$\tau_1 = -8{,}35$ $(\alpha = 1)$ або $\tau_1 = -4{,}04$ $(\alpha = 5)$;

$\tau_2 = -47{,}44$ $(\alpha = 1)$ або $\tau_2 = -17{,}83$ $(\alpha = 5)$.

2.1.5 Розширений тест Дікі-Фуллера

Для того щоб скористатись тестом ДФ, необхідно побудувати наступне рівняння регресії:

$$\Delta y(k) = a_0 + a_1 k + by(k-1) + \sum_{i=1}^{p} c_i \, \Delta y(k-i) + \varepsilon(k), \qquad (2.29)$$

де $a_0, a_1, b, c_i -$ невідомі коефіцієнти регресії.

Якщо всі коефіцієнти $c_i = 0, i = 1,2,..., p$, то рівнянням (2.29) можна скористатись для застосування тесту ДФ, інакше необхідно використати розширений тест ДФ. На практиці рекомендують застосовувати тест РДФ з кількістю затриманих у часі значень p меншою 10% числа спостережень, тобто $p < 0{,}1\,N$, де $N -$ довжина (потужність) часового ряду. При використанні тесту ДФ і РДФ важливо правильно задати структуру моделі, зокрема, необхідно визначити чи потрібно включати члени a_0 і $a_1 k$.

Пропонується наступне евристичне правило: якщо візуально з графіка не можна зробити висновок про наявність тренду, то в модель (2.29) необхідно включати тільки константу (перетин) a_0, навіть якщо значення коливаються навколо нуля. Якщо візуальний аналіз ряду свідчить про присутність тренду, то в модель (2.29) необхідно ввести a_0 і $a_1 k$.

За допомогою базової моделі (2.29) тестуються такі гіпотези:

$H_0 : b = 0$, або часовий ряд нестаціонарний: $\{y(k)\} \square I(\text{int})$, int > 0;

$H_1 : b < 0$, або часовий ряд стаціонарний: $\{y(k)\} \square I(0)$, int = 0.

Нульова гіпотеза відкидається, якщо отримана оцінка коефіцієнта $\hat{b} < 0$ та обчислена $\tau -$статистика Маккіннона (для тестування на наявність одиничного кореня) за абсолютною величиною більша за величину критичного значення цієї статистики при вибраному рівні значимості α.

Формально це можна записати так:

$$|\tau| = \left| \frac{\hat{b}}{SE_{\hat{b}}} \right| \geq |\tau_{крит}| \qquad (2.30)$$

на рівні значимості α; $SE_{\hat{b}} -$ стандартна похибка оцінки $\hat{b}$.

2.2 Моделі процесів з довгою пам’яттю

У фінансових програмах такі спостереження як (log-) повертаються часто некоррельованими, але не незалежними. На фінансовій мові це трактується як сильна залежність від нестабільності, зокрема в тому сенсі, що висока волатильність схильна до кластеризації. Це призводить до розвитку нелінійних моделей у тому сенсі, що умовна дисперсія залежить від минулого і, можливо, також і від самого часу. Для короткочасної мінливості волатильності існує розширена література, ініційована патріотичною роботою Енгла (1982) та Боллерслевом (1986), які відповідно представили моделі ARCH (p) та GARCH (p, q). Окрім прикладної роботи, існує величезна література, яка описує математичні властивості, такі як стаціонарність, поведінка хвоста, залежність, оцінка та обмеження теорем для GARCH та пов'язаних моделей. Проте, моделі GARCH (p, q) не можуть пояснити емпіричне спостереження, що часто залежність від мінливості досить сильна і довготривала, хоча цей процес, як і раніше, є стаціонарним. Отже, питання полягає в тому, щоб або розширити моделі GARCH, або визначити нові моделі, щоб включити довгострокову залежність. Першим природним продовженням є так званий ARCH (∞) процес. Загальна основа була введена в Робінсоні (1991a). Станція та властивості залежностей вивчались в Кокошці та Лейноні (2000 р.), Гіраїті, Кокошці та Лейпусі (2000 р.), Казакевичем і Лейпусом (2002 р., 2003 р.), Гіраїтом, Лейпусом і Сургалісом (2006 р.) Та Дук, Рууфом і Суліє 2008) серед інших. На перший погляд це розширення, здається, аналогічно модифікації моделей ARMA (p, q) до MA (∞) -процесів з несумміруючими вагами (див. Розділ 2.1.1.4).

Однак, як з'ясовується, стаціонарна послідовність ARCH (∞) з кінцевою дисперсією повинна мати сумарні ваги, і це виключає тривалу пам'ять. За аналогією з процесами IGARCH можна визначити моделі IARCH (∞) та FIGARCH

(Baillie, Bollerslev та Mikkelsen 1996), які обов'язково мають нескінченну дисперсію. Існування суто стаціонарного рішення було доведено в Douc, Roueff і Soulier (2008). Проте властивості залежності, включаючи інтерпретацію довгої пам'яті, незрозумілі.

Оскільки модель ARCH (∞) не може фіксувати довгу пам'ять у волатильності, альтернативою є так званий процес LARCH (∞), представлений Робінсоном (1991). Її стаціонарність та властивості залежностей вивчалися в Giraitis (2000b, 2003c, 2004), теоретичні оцінки та обмеження були розглянуті в Giraitis, Robinson і Surgailis (2000), Berkes and Horva th (2003), Beran (2006), Beran і Фенг (2007), Беран і Шунтеннер (2009). Крім того, Giraitis і Surgailis (2002) розглянули білінарні моделі, що складаються з комбінації довгої пам'яті в середньому з тривалим збереженням у волатильності, описаною структурою LARCH (∞). Оскільки процес умовного масштабування σt в моделях LARCH (∞) може бути негативним, Surgailis (2008) представив так званий процес LARCH + (∞), де гарантовано σt> 0. Цей процес також може захоплювати тяжку хворобу.

Вивчення властивостей процесів GARCH (p, q), ARCH (∞) або LARCH (∞) може бути математично досить вимогливим. Навпаки, встановлення властивостей існування, стаціонарності та властивостей залежностей, як правило, досить легко для моделей так званої "стохастичної мінливості". Першою моделлю цього типу є процес EGARCH, введений Нельсоном (1990) і поширюється на довгу установку пам'яті (під назвою FIE-GARCH) від Боллерслева та Міккельсена (1996). Незалежно, Брейт, Крато і де Ліма (1998) ввели дещо іншу стохастичну мінливість довгої пам'яті (також називається LMSV). Подальші дослідження можна знайти в Робінсоні та Заффароні (1997, 1998). Для стаціонарності та асимптотичних властивостей див., Наприклад, Гарвей (1998) і Сургайліс і Віано (2002), для розширень з важкими хвостами див. Девіс і Мікош (2001), Кулик і Сульє (2011, 2012a, 2012b). Для відгуків в економетричному контексті дивіться, наприклад, Байлі (1996) і Генрі і Заффароні (2003).

Щоб бути більш конкретним, ми починаємо з неформального визначення моделей волатильності. Слідом за Giraitis, Leipus і Surgailis (2006) поняття стохастичної мінливості зазвичай стоїть на моделях форми:

$$X_t = \sigma_t \varepsilon_t, \tag{2.31}$$

де ε_t (t ∈ Z) - випадкові величини з середньою нульовою та одиничною дисперсією, а σ_t - це (зазвичай позитивна) вимірювана функція колишніх значень ε_s, Xs (s ≤ t - 1) і, можливо, деяка додаткова, неперевірена інформація. Крім того, ε_s (s ≥ t) не залежить від ε_s , Xs (s ≤ t - 1). Це випливає з цього

$$E(X_t \mid \sigma_s, \varepsilon_s, \mathrm{s} \leq \mathrm{t} - 1) = 0 \tag{2.32}$$

також,

$$Var(X_t \mid \sigma_s, \varepsilon_s, \mathrm{s} \leq \mathrm{t} - 1) = {\sigma_t}^2 \tag{2.33}$$

Однак слід зазначити, що насправді немає стандартної термінології. Тобто, в контексті ціноутворення похідних «випадкова волатильність» часто відноситься до особливого випадку, коли послідовності σ_t (t ∈ Z) і ε_t (t ∈ Z) взаємно незалежні. Якщо це не так, то мова йде про "стохастичну мінливість з важелем".

Тепер ми детально обговоримо найважливіші моделі.

2.2.1 АРУГ (Авторегресія з умовною гетероскедастичністю)

Авторегресійна умовна гетероскедастичність АРУГ (англ. ARCH; AutoRegressive Conditional Heteroscedasticity). Дану модель використовують в економетриці для аналізу часових рядів (в першу чергу фінансових), у яких умовна дисперсія ряду залежить від минулих значень ряду, минулих значень цих дисперсій та інших факторів. АРУГ моделі призначені для «пояснення»

кластеризації волатильності на фінансових ринках, коли періоди високої волатильності тривають деякий час, змінюючись потім періодами низької волатильності, причому середню (довгострокову, безумовну) волатильність можна вважати відносно стабільною.

Авторегресійне умовно гетероскедастичне рівняння має вигляд:

$$\hat{\varepsilon}^2(k) = \alpha_0 + \alpha_1\hat{\varepsilon}^2(k-1) + \alpha_2\hat{\varepsilon}^2(k-2) + \cdots + \alpha_q\hat{\varepsilon}^2(k-q) + v(k), \qquad (2.34)$$

де $\hat{\varepsilon}^2(k)$ – квадрати оцінок залишків (похибок) моделі;

α_0 – коефіцієнт затримки;

$\alpha_1, \dots, \alpha_q$ – параметри;

$v(k)$ – процес білого шуму з нульовим середнім для адекватної моделі.

Залишки (збурення) $\varepsilon(k)$ можуть бути отримані на основі рівнянь регресії, авторегресії або авторегресії з ковзним середнім низького порядку.

Окрім рівняння типу (2.34) можна вибрати і складніші форми описання поведінки дисперсії. Наприклад, майже ніколи наперед невідомо як впливає збурення на процес – адитивно чи мультиплікативно. Тому його можна ввести в модель у мультиплікативній формі:

$$\varepsilon^2(k) = v^2(k)[\alpha_0 + \alpha_1\varepsilon^2(k-1)], \qquad (2.35)$$

де $v(k)$ – мультиплікативне збурення у формі білого шуму, причому $\{v(k)\}\sim(0,1)$, тобто воно має нульове середнє і одиничну дисперсію;

$\varepsilon(k-1)$ і $v(k)$ – статистично незалежні (некорельовані) величини.

Основним недоліком АРУГ є те, що $\alpha_i, i = 0, \dots q$ мають бути невід'ємними, щоб умовна дисперсія завжди була позитивною. Даний недолік дозволяє уникнути УАРУГ-моделі.

2.2.2 Узагальнений АРУГ (GARCH)

Узагальнена авторегресійна умовно гетероскедастична модель УАРУГ (англ. GARCH; Generalized Autoregressive Conditional Heteroscedastic) не має недоліку АРУГ. УАРУГ-модель передбачає, що на поточну зміну дисперсії впливають як попередні зміни показників, так і попередні оцінки дисперсії.

Розширення АРУГ моделі є описання умовної дисперсії як процесу АРКС. Нехай похибки описуються рівнянням

$$\varepsilon(k) = v(k)\sqrt{h(k)}, \tag{2.36}$$

де $\sigma_v^2 = 1$;

$h(k)$ – умовна дисперсія визначається за виразом

$$h(k) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i), \tag{2.37}$$

де p – кількість попередніх оцінок, які впливають на поточне значення;

β_i – вагові коефіцієнти, які відображають степінь впливу попередніх оцінок на поточне значення.

УАРУГ(p, q) складається з двох компонент – авторегресії та ковзного середнього відносно дисперсії гетероскедастичного процесу.

По відношенню до узагальнених умовно гетероскедастичних процесів не узгоджено визначення стаціонарності, тобто сильно стаціонарний процес УАРУГ не завжди буде слабо стаціонарним. Таким чином, виникає проблема визначення стаціонарності таких процесів.

2.2.3 АРУГ (∞) ПРОЦЕСИ

Використовуючи загальні коефіцієнти bj умовної дисперсії призводить до наступного визначення.

$$X_t = \sigma_t \varepsilon_t, \tag{2.38}$$

$$\sigma_t{}^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}{}^2, \tag{2.39}$$

Зазвичай також передбачається, що перші два моменти εt є кінцевими і σε2 = var (εt) = 1. Причиною останнього припущення є ідентифікація, оскільки статистично параметр $\sigma_\varepsilon{}^2$ не відрізняється від b. Більш загальне визначення дано Робінсоном (1991).

2.2.4 ЕКСПОНЕНЦІЙНИЙ УАРУГ (EGARCH)

Експоненційна узагальнена авторегресійна умовно гетероскедастична модель ЕУАРУГ (англ. EGARCH; Exponential Generalized Autoregressive Conditional Heteroscedastic) не має недоліків АРУГ та УАРУГ моделей. В цій моделі логарифм умовної дисперсії визначається за допомогою функції нормованих похибок g(y):

$$\log[h(k)] = c_0 + \sum_{i=1}^{\infty} c_i g[y(k-i)], \tag{2.40}$$

$$g(y) = \alpha y(k) + \beta[|y(k)| - E|y(k)|], \tag{2.41}$$

де $E[g(y)] = 0$;

α, β – параметри моделі;

$y(k)$ – основна (залежна) змінна, що моделюється.

– Узагальнена білінійна модель

Дана модель широко застосовується і має зручну структуру:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j v(k-j) + \sum_{i=1}^{m}\sum_{j=1}^{s} c_{i,j} y(k-i)v(k-j) + \varepsilon(k), \qquad (2.42)$$

де p, q, m і s є позитивними числами, що відображають модельний порядок.

– Лінійна комбінація лінійних та нелінійних компонентів

Дуже часто моделювання нелінійних процесів ґрунтується на лінійній комбінації лінійних та нелінійних компонентів:

$$y(k) = \beta^T z(k) + \sum_{i=1}^{p} \alpha_i \varphi_i(\theta_i^T z(k)) + \varepsilon(k), \qquad (2.43)$$

де $z(k)$ – вектор значень затримки часу залежної змінної $y(k)$, а також попередні та поточні значення незалежних пояснювальних змінних $\mathbf{x}(k)$ з відповідним зміщенням часу;

φ_i – набір (лінійних та нелінійних) функцій, які включають наступні компоненти: функцію потужності $\varphi_i(x) = x^i$, тригонометричні функції $\varphi_i(x) = \sin x$ або $\varphi_i(x) = \cos x$ та ін.

2.2.5 Модель лінійний АРУГ (LARCH ((∞))

Як згадувалося вище, процеси стаціонарного ARCH (∞) другого порядку не можуть фіксувати довгу пам'ять у волатильності. Робінзон (1991) представив так званий лінійний процес ARCH (LARCH), який визначається як

$$X_t = \sigma_t \varepsilon_t \tag{2.44}$$

і

$$\sigma_t = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}, \tag{2.45}$$

де ε_t нульові середні випадкові величини з $\sigma_t{}^2 = E(\varepsilon_t{}^2) = 1$. Модель форми (2.30) і, отже, $E(X_t \mid \sigma_s, \varepsilon_s, \mathrm{s} \leq \mathrm{t}) = 0$. Крім того, X_t – мартингал. Суттєвою модифікацією порівняно з ARCH (∞) -процесами є що σ замість σ^2 виражається як лінійна функція X (замість X^2). Суворе ставлення до імовірнісних аспектів, таких як стаціонарність та припущення до моменту. В роботах Giraitis, Robinson і Surgailis (2000), Giraitis, Leipus, Robinson та Surgailis (2004) було надано таке уявлення про імовірнісні аспекти, як стаціонарність та припущення про момент. Як ми побачимо нижче, умовна дисперсія $\sigma_t{}^2$в LARCH (∞).

Модель може мати довгу пам'ять, яка на відміну від ARCH (∞) моделей. З іншого боку, σt може стати негативним, так що воно може бути складніше інтерпретувати як волатильність. Перше питання, яке потрібно вирішити, полягає в тому, чи існує стаціонарне рішення. Будуть використані наступні позначення:

$$\|b\|_p = \left(\sum_{j=1}^{\infty} |b_j|^p \right)^{\frac{1}{p}},$$
$$\mu_p = E\left[\varepsilon_t^p\right], |\mu|_p = E\left[|\varepsilon_t|^p\right] \tag{2.46}$$

де p ∈ N. За повторної ітерації (2.46) варіантом для рішення може бути

$$\sigma_t = b_0 + \sum_{k=1}^{\infty} \sum_{j1,\dots,jk=1}^{\infty} b_{j_1} \dots b_{j_k} \varepsilon_{j_1} \dots \varepsilon_{j_k}, \quad (2.47)$$

2.3 Критерії для аналізу адекватності моделей ННП

Формально адекватність визначають за допомогою ряду статистичних величин. Наприклад, дуже часто використовують середньо-квадратичну похибку моделі (СКП), яка обчислюється за формулою:

$$СКП(x_s, x_m) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [x_s(k) - x_m(k)]^2}, \quad (2.48)$$

де $x_s(k)$ – вимір вихідного сигналу об'єкта в момент k;

$x_m(k)$ – оцінка вихідного сигналу об'єкта, отримана по оціненій моделі.

Для лінійних моделей запропоновано кілька статистичних параметрів, що використовуються при оцінюванні адекватності, які будуть розглянуті нижче. Використання одного параметра для визначення ступеня адекватності моделі є некоректним підходом, оскільки оцінки параметрів – це випадкові величини, а тому збільшення числа критеріїв адекватності сприяє підвищенню ймовірності вибору адекватної моделі.

Рівняння моделі повинні мати розв'язок, тобто бажано мати аналітичний або, якщо це неможливо, то чисельний розв'язок.

Одним із принципів, яких необхідно дотримуватись при побудові моделі є наступний: "в моделі не повинно бути нічого зайвого крім необхідного". Звичайно, що дотримуватись цього принципу досить непросто, і на практиці буває так, що модель дійсно має надзвичайно складну структуру, що також може

бути оправдано необхідністю досягнення високого ступеня її адекватності процесу. Це особливо стосується нелінійних процесів. Але при побудові лінійних моделей у вигляді авторегресії чи авторегресії з ковзним середнім достатньо побудувати модель, статистичні характеристики якої співпадають з статистичними характеристиками часового ряду, на основі якого вона оцінюється. Такі спрощені моделі виявляються цілком придатними для прогнозування та керування процесами. Загалом питання складності моделі вирішується в кожному випадку окремо.

Модель повинна буди достатньо універсальною, щоб її можна було застосувати до описання класу однотипових процесів або до описання функціонування процесу в різних умовах.

Наприклад, для описання моторної функції людини (реакція на зовнішні збуджуючі сигнали) застосовують звичайне диференціальне рівняння другого порядку, яке представляють у вигляді функції передачі такого ж порядку:

$$W(s) = \frac{K\, e^{-\tau s}}{(1 - T_1 s)(1 - T_2 s)}, \tag{2.49}$$

де K – статичний коефіцієнт передачі об’єкта;

τ – час запізнення по входу, який в середньому дорівнює для людини 300-350 мс;

T_1, T_2 – постійні часу. Така передаточна функція може використовуватись, наприклад, для описання реакції людини на зовнішні відео- або аудіосигнали, що поступають через систему візуального сприйняття чи аудіосистему (поширений приклад – водіння автомобіля чи іншої машини). Значення параметрів моделі можуть бути різними для різних людей, але структура моделі залишається

незмінною. Таким чином, наведена модель описує широкий клас біологічних систем і цілком відповідає умові універсальності.

При моделюванні технічних систем широко застосовують ланки першого і другого порядку, що відповідають звичайним диференціальним рівнянням таких же порядків. На основі таких простих ланок можна побудувати моделі будь-якої складності. Дуже поширений в техніці та екології клас систем з розподіленими параметрами. Наприклад, процес розповсюдження домішок в атмосфері та водному середовищі, механічні коливання сонячних батарей та антен супутників, крила літака, локомотива з вагонами на залізниці, автомобіля з причепом і багато інших. Динаміку таких систем описують диференціальними рівняннями з частковими похідними.

Вимога робастності (robust – сильний, міцний). Робастність означає, що модель повинна давати прийнятний прогноз вихідної змінної не тільки на тому відрізку часового ряду, на основі якого вона побудована, але і на будь-якому іншому відрізку, що відповідає вибраному режиму роботи. Робастність може розглядатись також як стійкість моделі по відношенню до збурень, похибок та пропусків вимірів. Вимога робастності є особливо критичною для систем, що працюють в реальному часі, оскільки нестійка модель може стати причиною створення аварійної ситуації.

Вимога адаптивності. Ця вимога означає, що хоча б частину параметрів моделі можна уточнювати по мірі надходження нових вимірів від об’єкта. Ця вимога є обов’язковою при побудові моделей нестаціонарних систем, тобто систем, параметри яких є функціями часу. Системи керування, побудовані для нестаціонарних процесів, називають адаптивними. Такі системи є досить складними з точки зору аналізу збіжності оцінок параметрів та похибок керування, а тому при проектуванні адаптивних систем необхідно особливу увагу приділяти задачам достатнього збудження процесу та вибору методу оцінювання параметрів.

Коли аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика складається з наступних кроків:

а) *Візуальне дослідження графіка похибок* моделі

$$e(k) = y(k) - \hat{y}(k), \tag{2.50}$$

де $\hat{y}(k)$ - оцінка змінної, отримана за допомогою побудованої моделі.

На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка приймає великі значення (тобто довгих інтервалів суттєвої неадекватності). У випадку застосування рекурсивних методів оцінювання найбільші похибки будуть в перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації про процес.

б) *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ та ЧАКФ для ряду $\{e(k)\}$ і за допомогою Q - статистики визначити ступінь корельованості (наприклад, Q - статистика вважається несуттєвою до рівня 10%).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою:

$$DW = 2 - 2\rho, \tag{2.51}$$

де $\rho = E[e(k)e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки;

σ_e^2 – дисперсія послідовності похибок $\{e(k)\}$.

Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Отримати формулу $DW = 2 - 2\rho$ можна досить просто. Автори цієї статистики (Durbin, Watson) запропонували скористатись для перевірки корельованості похибок моделі наступним виразом:

$$DW = \frac{\sum_{k=2}^{N}\left[e(k)-e(k-1)\right]^2}{\sum_{k=1}^{N}e^2(k)} = \frac{\sum_{k=2}^{N}\left[e(k)-e(k-1)\right]\left[e(k)-e(k-1)\right]}{\sum_{k=1}^{N}e^2(k)}, \tag{2.52}$$

тобто, DW можна, в деякій мірі, трактувати як коефіцієнт автокореляції для (перших різниць) приростів похибок.

Розкриваючи квадрат різниці в чисельнику, отримаємо:

$$DW = \frac{\sum_{k=2}^{N}e^2(k)}{\sum_{k=1}^{N}e^2(k)} + \frac{\sum_{k=2}^{N}e^2(k-1)}{\sum_{k=1}^{N}e^2(k)} - 2\,\frac{\sum_{k=2}^{N}e(k)e(k-1)}{\sum_{k=1}^{N}e^2(k)}\ , \tag{2.53}$$

де $\dfrac{\sum_{k=2}^{N}e^2(k)}{\sum_{k=1}^{N}e^2(k)} \approx 1;$ $\dfrac{\sum_{k=2}^{N}e^2(k-1)}{\sum_{k=1}^{N}e^2(k-1)} \approx 1;$ а $\dfrac{\sum_{k=2}^{N}e(k)e(k-1)}{\sum_{k=1}^{N}e^2(k-1)} = \rho\,.$

Тому можна записати, що $DW = 2 - 2\rho\,.$

в) Для лінійної моделі 2-3 порядку *оцінки параметрів повинні збігатися до усталених значень* після 30-40 (не більше) ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує вказані числа, то це свідчить про те, що процес може бути нестаціонарним.

г) Перевірка значимості параметрів моделі. *Статистика Стьюдента або* $t-$*статистика* (випадкова величина, що має $t-$розподіл), яка використовується для визначення значимості оцінки кожного коефіцієнта в статистичному смислі, визначається за виразом:

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}}, \tag{2.54}$$

де $\hat{a}$ – оцінка коефіцієнта моделі;

a^0 – нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки;

$SE_{\hat{a}}$ – стандартна похибка оцінки.

За нуль-гіпотезу щодо значимості оцінки можна висувати будь-яку: що коефіцієнт значимий, тобто, $(H_0 : a^0 \neq 0)$ або незначимий $(H_0 : a^0 = 0)$. Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. В даному випадку бажаним результатом є значимість коефіцієнтів математичної моделі. Таким чином, необхідно висувати нульову гіпотезу, що коефіцієнт незначимий. Це дає можливість коректно підійти до визначення значимості оцінок коефіцієнтів та дещо спростити розрахунки.

Для того щоб встановити, чи є оцінка коефіцієнта значимою, необхідно знати довжину вибірки даних N (потужність вибірки); число ступенів свободи $f = N - n$, де $n-$ число коефіцієнтів моделі, які оцінюються на основі ряду даних, і вибрати рівень значимості $\alpha = 1\%$ або $\alpha = 5\%$ або $\alpha = 10\%$ (для цих значень існують розраховані таблиці для критичних значень $t-$статистики). Фактично, рівень значимості означає ймовірність припуститись *помилки першого роду* при перевірці гіпотези. Згадаємо, що

$$\alpha = p\{X \in G / \omega \,|\, H_0\} = \int_{n-m(G/\omega)} L_{H_o}(X)\, dx, \tag{2.55}$$

де $X = [x_1, ..., x_n] \in R^n -$ вся вибірка, яка розбивається на дві множини, що перетинаються: ω і G/ω ($\omega -$ область прийняття нуль-гіпотези);

$G/\omega -$ критична область: якщо $X \in G/\omega$, то H_0 відхиляється; $L_{H_0}(X) -$ закон розподілу X. Помилка першого роду означає відхилення вірної гіпотези.

Користуючись значеннями N, f і α, з таблиць для $t-$ розподілу знаходять критичне значення $t-$статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо $-t_{кр} < t < t_{кр}$ або $|t| < |t_{кр}|$, то нуль-гіпотеза щодо незначимості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більшим є значення $t-$статистики для оцінки коефіцієнта, тим імовірніше, що цей коефіцієнт є значимим.

Загалом послідовність дій при перевірці значимості оцінок коефіцієнтів побудованої моделі можна сформулювати так:

– сформулювати нуль-гіпотезу щодо значимості коефіцієнта;

– обчислити значення $t-$ статистики для кожного коефіцієнта регресії (це робить кожний пакет для математичного моделювання);

– за допомогою значень N, f і α знайти із таблиць для $t-$статистики її критичне значення;

– перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр} < t < t_{кр}$).

д) *Коефіцієнт множинної детермінації* R^2, який обчислюється так:

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)} = 1 - \frac{SSE}{SST}, \tag{2.56}$$

де $\text{var}(\hat{y})$ – дисперсія залежної змінної, оціненої за допомогою побудованої моделі;

$\text{var}(y)$ – дисперсія вимірів залежної змінної;

$SSE = \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2$ – сума квадратів похибок (залишків) моделі (*sum of squared errors*);

$SST = \sum_{k=1}^{N}[y(k) - \bar{y}]^2 -$ загальна сума квадратів (*total sum of squares*);

$\bar{y} -$ середнє значення;

$SST = SSE + SSR,$

де $SSR = \sum_{k=1}^{N}[\hat{y}(k) - \bar{y}]^2 -$ загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є $R^2 = 1$, тобто, коли дисперсії вимірів змінної, та цієї ж змінної, оціненої за рівнянням, збігаються. Цей параметр можна трактувати, також, як міру інформативності моделі, якщо вибрати за міру інформативності дисперсію. Таким чином, R^2 показує рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

е) *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто,

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N}[\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}} \tag{2.57}$$

у порівнянні з усіма іншими моделями.

є) Для оцінки адекватності моделі також використовують *інформаційний критерій Акайке*

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n \tag{2.58}$$

та *критерій Байєса-Шварца*

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N), \tag{2.59}$$

де $n = p + q + 1$ - число параметрів моделі, які оцінюються за допомогою статистичних даних (p - число параметрів авторегресійної частини моделі;

q - число параметрів ковзного середнього; 1 з'являється тоді, коли оцінюється зміщення (або *перетин*), тобто a_0).

Критерії Акайке і Байєса-Шварца містять в правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої критерії приймають найменші значення. Введення нового регресора приводить до збільшення критерію (при цьому збільшується n), але, разом з тим, зменшується сума квадратів похибок і критерій в цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для довгих виборок кращі у критерія Байєса-Шварца, тобто, його рекомендують застосовувати при відносно великих значеннях N (N>100).

ж) Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують F - *статистику Фішера*, яка пропорціональна відношенню:

$$F \sim \frac{R^2}{1 - R^2}, \tag{2.60}$$

а для множинної (багатофакторної) регресії вона визначається за виразом

$$F = \frac{R^2}{1 - R^2} \cdot \frac{(N - p - 1)}{p}, \tag{2.61}$$

де, як і раніше, N – число значень ряду;

p – число параметрів моделі без врахування перетину (константи).

Таким чином, якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування $F-$ статистики такий же, як і $t-$статистики. Нуль-гіпотезою є в даному випадку припущення про те, що модель неадекватна в цілому, тобто, $H_0 : a_1 = a_2 = ... = a_p = 0$ проти альтернативної гіпотези .

H_1: хоча б одне значення a_i відмінне від нуля в статистичном смислі.

Значення $F_{крит}$ знаходять із таблиць для $F-$ розподілу. Послідовність застосування цієї статистики можна представити наступним чином:

1. Сформулювати нуль-гіпотезу щодо адекватності моделі в цілому. Наприклад, H_0: *модель неадекватна в цілому* (або $H_0 : a_1 = a_2 = ... = a_p = 0$).
2. Розрахувати значення F для оціненої моделі (як правило, воно розраховується всіма пакетами статистичної обробки даних).
3. Задати рівень значимості $\alpha = 1\%$ або $\alpha = 5\%$ або $\alpha = 10\%$.
4. Користуючись значеннями N, f і α, знайти критичне значення $F_{крит}$ знаходять із таблиць для $F-$ розподілу при $(p, N-p-1)$ степенях свободи.
5. Перевірити нуль-гіпотезу: якщо $F > F_{крит}$, то нуль-гіпотеза щодо неадкватності моделі в цілому відкидається на вибраному рівні значимості.

б) Критерії адекватності нелінійних моделей

Статистика Хосмера-Лемешоу (H-L, HL, Hosmer-Lemeshow). Для розрахунку даної статистики вибірка розбивається на кілька підвибірок, по кожній з яких визначаються — фактична частка даних зі значенням залежної змінної 1, тобто фактично середнє значення залежної змінної за підвибіркою

$$p_j = \bar{y_j} = \sum_{i=1}^{n_j} y_{ij} / n_j. \qquad (2.62)$$

і передбачена середня вірогідність в підруппі:

$$\overline{\widehat{p_j}} = \sum_{i=1}^{n_j} \widehat{p_{ij}}/n_j. \tag{2.63}$$

Тоді значення статистика Хосмера-Лемешоу визначається по формулі:

$$HL = \sum_{j=1}^{j} \frac{n_j(p_j - \overline{\widehat{p_j}})}{\overline{\widehat{p_j}}(1 - \overline{\widehat{p_j}})}. \tag{2.64}$$

Точний розподіл даної статистики невідомий, однак автори методом симуляцій встановили, що він апроксимується розподілом $\chi^2(J-2)$.

в) Критерії якості оцінок прогнозів

Важливим етапом прогнозування є верифікація прогнозів, тобто оцінювання їх точності та обґрунтованості. На етапі верифікації використовують сукупність критеріїв, способів і процедур" які дають можливість оцінити якість прогнозу.

Найбільш поширене ретроспективне оцінювання прогнозу, тобто оцінювання прогнозу для минулого часу.

Для цього вихідна інформація поділяється на дві частини, одна з яких охоплює більш ранні дані, а інша - більш пізні. За допомогою даних першої групи (ретроспекції) оцінюються параметри моделі прогнозу, а дані другої групи розглядаються як фактичні дані прогнозованого показника. Отримана ретроспективно помилка прогнозу певною мірою характеризує точність застосовуваної методики прогнозування.

Усі показники, що використовуються для аналізу якості прогнозу, можна розділити на три групи: абсолютні, порівняльні та якісні.

До абсолютних відносять показники, які дають змогу кількісно визначити величину помилки прогнозу в одиницях виміру прогнозованого об'єкта або у відсотках. Це середньоквадратична помилка σ_t, абсолютна помилка $\Delta_{пр}$, середня абсолютна помилка $\overline{\Delta_{пр}}$, відносна помилка ε та середня відносна помилка прогнозу$\bar{\varepsilon}$.

Абсолютна помилка прогнозу може бути визначена як різниця між фактичним значенням (у) і прогнозом (у*):

$$\Delta_{\text{пр}} = y_t - y^*. \tag{2.65}$$

Середнє абсолютне значення помилки становитиме:

$$\overline{\Delta_{\text{пр}}} = \frac{\sum_{t=1}^{n} |y_t - y_t^*|}{n}. \tag{2.66}$$

Середньоквадратична помилка прогнозу розраховується за формулою:

$$\sigma_t = \sqrt{\frac{\sum_{t=1}^{n} (y_t - y_t^*)^2}{n}}. \tag{2.67}$$

Слід зазначити, що для великого класу статистичних розподілів існує зв'язок середнього абсолютного відхилення σ_t зі стандартним відхиленням $\overline{\Delta_{\text{пр}}}$, що можна представити в такому вигляді:

$$\sigma_t = 1.25\overline{\Delta_{\text{пр}}}. \tag{2.68}$$

Недоліком розглянутих показників є те, що значення цих характеристик істотно залежить від масштабу виміру рівнів досліджуваних явищ.

Відносна помилка прогнозу ε може бути виражена у відсотках щодо фактичних значень показника в такий спосіб:

$$\varepsilon = \frac{y_t - y_t^*}{y_t} * 100. \tag{2.69}$$

а середня відносна помилка (помилка апроксимації) розраховується як

$$\bar{\varepsilon} = \frac{\sum_{t=1}^{n} \frac{|y_t - y_t^*|}{y_t} * 100}{n}. \quad (2.70)$$

Цей показник, як правило, використовується при порівнянні точності прогнозів різнорідних об'єктів прогнозування. Типові значення єпр для середньострокових прогнозів та їх інтерпретації наведено в табл. 2.1.

Таблиця 2.1 - Дані прогнозування та інтерпретації

ε	Інтерпретація
<10	Висока точність
10-20	Хороша точність
20-50	Задовільна точність
>50	Незадовільна точність

Середня абсолютна і середньоквадратична помилки фіксують середнє значення помилки на кожному стані прогнозу без урахування цієї помилки. Середня помилка дає змогу визначити, який вид помилки є найбільш типовим - недооцінка чи переоцінка прогнозованого показника. Необхідно мати на увазі, що $\Delta_{пр}$ і $\overline{\Delta_{пр}}$, дорівнюють нулю тільки тоді, коли y_t - y* для кожного t, тобто у випадку досконалого прогнозу. Аналогічне твердження несправедливе для абсолютної помилки , оскільки тут може мати місце взаємопогашення помилок. Для розрахунку цих показників можуть бути використані як абсолютні величини змінних, так і їхні прирости.

Порівняльні показники точності прогнозу ґрунтуються на порівнянні помилки розглянутого прогнозу з еталонними прогнозами певного виду.

Один із типів таких показників К може бути в загальному вигляді поданий так:

$$К = \sqrt{\frac{\sum_{t=1}^{n}(p_t - y_t)}{\sum_{t=1}^{n}(p_t^* - y_t)}}. \quad (2.71)$$

де p_t^* - прогнозоване значення величини еталонного прогнозу.

Як еталонний прогноз може бути обрана проста екстраполяція, простий темп приросту і т. ін.

Окремим випадком показників такого типу є коефіцієнт невідповідності, у якому $p_t^* = 0$ для всіх типів p_t:

$$KH = \sqrt{\frac{\sum_{t=1}^{n}(y_t^* - y_t)^2}{\sum_{t=1}^{n} y_t^2}}. \quad (2.72)$$

Можна побудувати різні модифікації коефіцієнта невідповідності. Розглянемо деякі з них.

1. Коефіцієнт невідповідності (KH_1 обчислюється як відношення середньоквадратичної помилки прогнозу до тієї самої помилки, що мала б місце, якщо брати як прогноз для кожного року середнє значення змінної за весь період:

$$KH_1 = \sqrt{\frac{\sum_{t=1}^{n}(y_t^* - y_t)^2}{\sum_{t=1}^{n}(\bar{y} - y_t)^2}}. \quad (2.73)$$

Якщо $KH_1 > 1$, то прогноз на рівні середнього значення дав би кращий результат, ніж отриманий прогноз.

2. Коефіцієнт розбіжності V становить відношення середньоквадратичної помилки прогнозу до тієї самої помилки, що мала б місце, якщо брати як прогноз для кожного року значення, вирівняне по аналітичному тренду, тобто

$$V = \sqrt{\frac{\sum_{t=1}^{n}(y_t^* - y_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}}. \quad (2.74)$$

Якщо V > 1, то прогноз методом простої екстраполяції дає кращий результат.

До порівняльних показників варто віднести і коефіцієнт кореляції між прогнозованими і фактичними значеннями змінної.

Одним із недоліків використання коефіцієнта кореляції як вимірника точності прогнозів є те, що повна позитивна кореляція лише вказує на наявність лінійної залежності між низкою прогнозних і фактичних величин. Унаслідок цього коефіцієнт кореляції найбільш придатний для аналізу прогнозів змінних, що циклічно розвиваються.

Якісні показники точності прогнозу дають змогу провести аналіз видів помилок прогнозу, поділити їх на складові. Особливо такий аналіз є важливий для змінних, що циклічно змінюються, коли необхідно прогнозувати не лише загальний напрямок розвитку, але і поворотні точки циклу.

Одним із методів такого аналізу є діаграма "прогноз - реалізація". Сутність методу полягає в побудові точкових прогнозів у координатах, у яких на одній осі відкладається реальне значення змінної, на іншій її прогнозоване значення (рис. 1.4).

Використання діаграми дає змогу змістовно оцінити якість різних прогнозів, розрахувати коефіцієнти, що аналізують якість прогнозування поворотних точок, виділити найбільш типові помилки (недооцінки або переоцінки змін). Для аналізу більш загальних видів помилок прогнозів може бути використана їх класифікація за джерелами виникнення.

Рисунок 2.2 - Діаграма "прогноз - реалізація"

2.4 Методологія DATA-MINING

Методологія Data Mining грунтується на підході побудови різноманітних моделей, які забезпечують при прийнятті рішень розв'язання задач класифікації, кластеризації, регресії, прогнозування, асоціації, виявлення послідовностей. При цьому дана методологія забезпечує реалізацію багаторівневих ієрархічних систем деталізації й узагальнення даних, одержання будь-яких їх проекцій, що дозволяють розгледіти як дрібні деталі, так і побачити картину в цілому. Приведемо схему яка описує методологію:

Рисунок 2.3 – Методологія data-mining

З схеми бачимо, що перед тим як застосовувати підходи і методи data-mining ми повині зробити попередню обробку даних(підготувати дані), потім оцінити параметри моделі, вибрати структуру моделі і протестувати її використовуючи критерії оцінки та зробити оцінку нашого прогнозу.

Висновки до розділу

В цьому розділі ми навели класифікацію моделей:

- Лінійні та нелінійні моделі
- Моделі процесів з довгою пам’яттю (АРУГ, УАРУГ, Е-УАРУГ моделі)

Кожні з цих моделей показуют себе з кращої сторони на деяких типах даних, які не дуже гарно описуються іншими моделями. Це справедливо для складних економічних процесів, які погано описуються стандартними моделями АР і АРКС.

Описали критерії адекватності моделей і далі будемо аналізувати їх на прикладі реальних даних .

Також ми детально розглянули методологію data-mining. Вхідну базу даних ми спочатку повинні обробити і підготувати для використання. Далі обираємо моделі, які будемо використовувати для опису наших даних і критерії адекватності. Це необхідно для тестування отриманих моделей. І у результаті оцінюємо прогноз та перевіряємо чи вся база даних була проаналізована.

РОЗДІЛ 3 РОЗРОБКА СППР ДЛЯ ВИКОНАННЯ ОБЧИСЛЮВАЛЬНИХ ЕКСПЕРИМЕНТІВ

3.1 Архітектура СППР

Система підтримки прийняття рішень або СППР — це комп'ютерна система, яка шляхом збору і аналізу великої кількості інформації може впливати на процес ухвалення рішень організаційного плану в бізнесі чи підприємництві. Інтерактивні системи дозволяють керівникам отримати корисну інформацію з першоджерел, проаналізувати її, а також виявити існуючі бізнес-моделі для вирішення певних завдань. Реалізована СППР є найбільш простою з погляду архітектури, тому її впровадження буде доцільним в організаціях, що не ставлять перед собою глобальних завдань і що мають невисокий рівень розвитку інформаційних технологій.

Архітектура створеної СППР налічує наступні рівні:

- завантаження і обробка даних;
- аналіз даних;
- побудова та вибір кращої моделі;
- прогнозування.

Розглянемо кожен із перерахованих рівнів:

- перший рівень надає можливість завантаження даних шляхом імпорту із текстового файлу чи ручним вводом. Після чого можливе перетворення даних, з метою усунення їх надлишковості, та підготовка даних до аналізу;

- другий рівень забезпечує можливість візуальної оцінки даних, проведення статистичного та кореляційного аналізу.

- третій рівень надає засоби для побудови моделей авто регресії ковзкого середнього. Надає можливість оцінки параметрів якості моделі для вибору кращої.

- четвертий рівень реалізує динамічне та статистичне прогнозування на базі створеної АРКС моделі.

Рисунок 3.1 - Рівні архітектури створеної СППР

3.2 Вибір інструментальної платформи для реалізації системи

Дослідження і всі розрахунки були проведені на платформі Eviews. Було використано багато бібліотек, додаткова бібліотека, що надає доступ до широкого кола статистичних методів прогнозування, в тому числі має запрограмований метод АРУГ, УАРУГ, Е-УАРУГ моделей.

Дане програмне забезпечення обране за декількома причинами:

- Високий рівень пристосованості до вирішення статистичних проблем і проблем прогнозування.
- Наявність великої кількості розроблених бібліотек і модулів, що часто дозволяє позбутися необхідності «винаходити велосипед».

- Дана платформа є інтерпретатором, що дозволяє досягти високої швидкодії.

Масив даних являє собою інформацію про значення закриття цін на фондому ринку по компаніям Apple, Google, Amazon, ціна на золото за 3 роки і температуру повітря в Нью Йорку за 3 роки.

На рисунку 3.1 зображено розподіл змінної Appl.

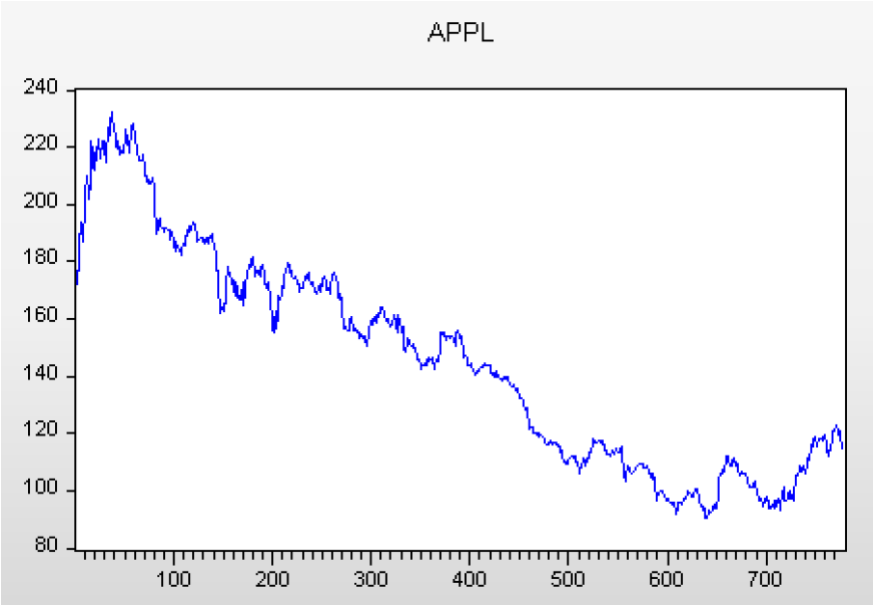

Рисунок. 3.1 - розподіл змінної Apple

3.3 Результати обчислювальних експериментів

3.3.1 Вікно програми

На рисунку 3.2 зображено вікно програми для роботи з курсами акцій компанії Apple.

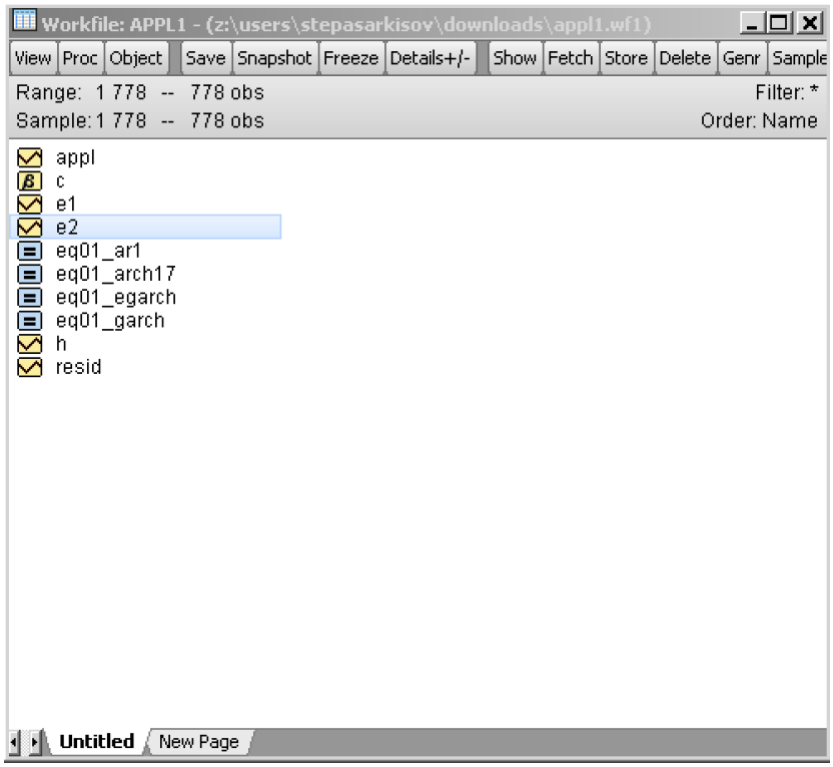

Рисунок 3.2 - Головний екран програми

Програма працює за допомогою натискання необхідних кнопок для виконування операцій.

Взаємодія відбувається поступово – кнопки розташовані зверху.

Основна структура роботи програми:

- Завантаження даних.
- Вибір моделі.
- Тренування моделі на основі завантажених даних.
- Використання моделі для прогнозування цільової змінної на текстовому наборі даних.

Далі ці етапи будуть розглянуті детальніше.

3.3.2 Робота програми

Розглянемо роботу програми на основі даних appl.txt (ціни акцій компанії Apple на біржі за 3 роки).

На рисунку 3.3 зображено вікно з імпортованими даними.

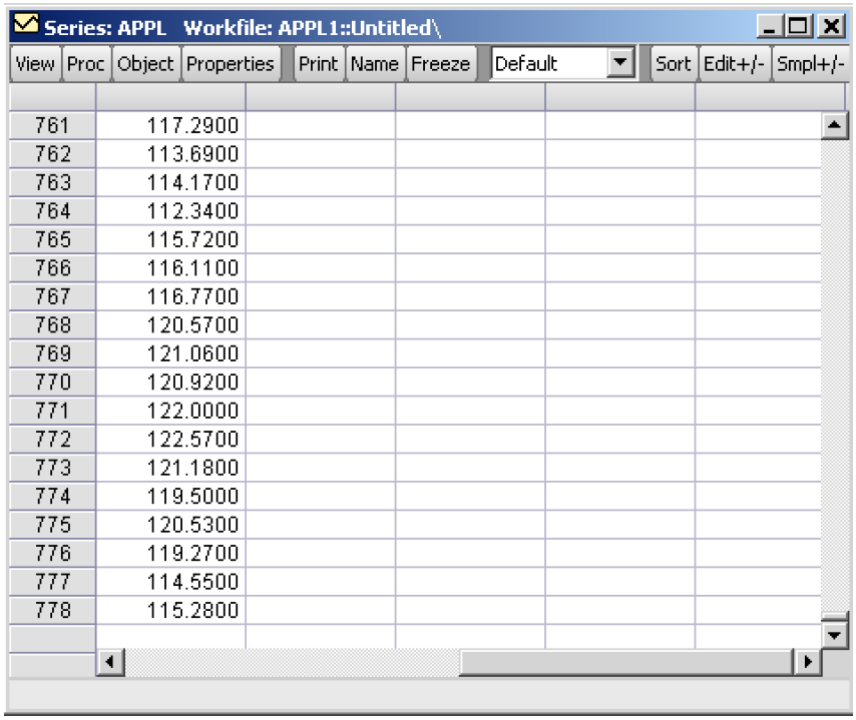

Рисунок 3.3 – Результати роботи програми після імпорту даних.

Далі розрахуємо зміну e1 – залишки регресійної моделі.

На рисунку 3.4 зображено вікно програми зміної e1.

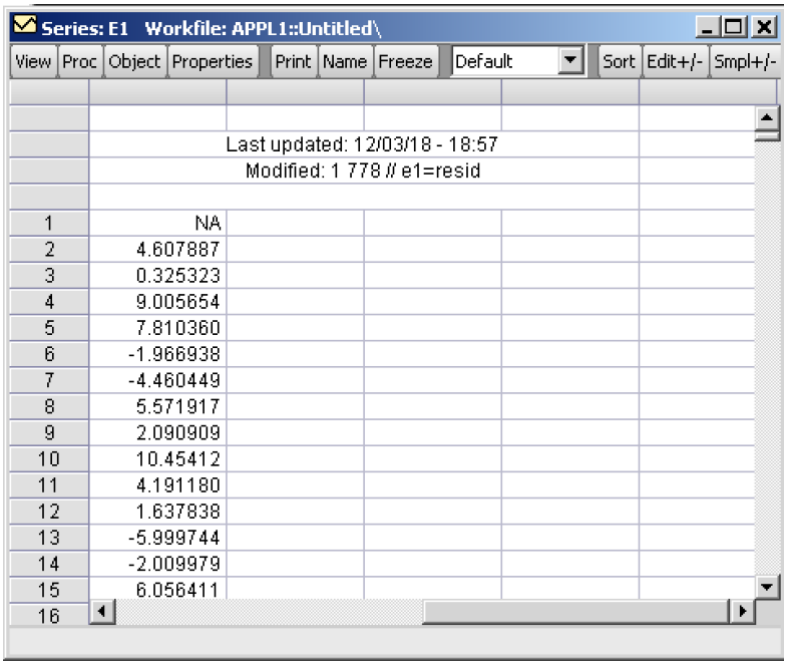

Рисунок 3.4 – Зміна e1

За аналогією розраховуємо зміну e2 – квадрат залишків і для розрахунку моделей подивимось на АКФ та ЧАКФ.

На рисунку 3.5 зображено авто-кореляційну та частково-кореляційну функції .

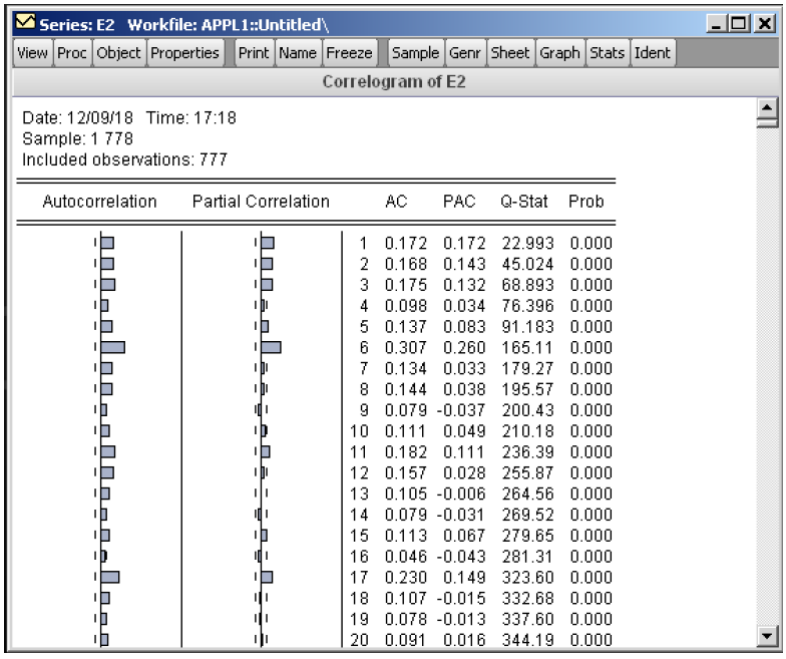

Рисунок 3.5 – Часткова автокореляція лагів

Далі для розрахунків обчислимо функцію H. Для цього використаємо програму, яка написана власноруч.

На рисунку 3.6

```
series h=NA
'sample cycle_sample @all
for !i=2 to 778
  smpl @first @first+!i-1
  h(!i)=@stdev(e2)
next !i
```

Рисунок 3.6 – Програма на Eviews для розрахунку H

На рисунку 3.7 зображено АРУГ модель для лагів 1, 2, 3, 5, 6, 11, 17.

Equation: EQ01_ARCH17 Workfile: APPL1::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

E2=C(1)+C(2)*E2(-1)+C(3)*E2(-2)+C(4)*E2(-3)+C(5)*E2(-5)+C(6)*E2(-6)
+C(7)*E2(-11)+C(8)*E2(-17)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1.993814	0.423638	4.706406	0.0000
C(2)	0.109124	0.035035	3.114679	0.0019
C(3)	0.107749	0.034950	3.082950	0.0021
C(4)	0.026657	0.027774	0.959776	0.3375
C(5)	-0.013724	0.028198	-0.486694	0.6266
C(6)	0.042652	0.027483	1.551932	0.1211
C(7)	-0.005440	0.027728	-0.196206	0.8445
C(8)	0.208250	0.026535	7.848257	0.0000

R-squared	0.136629	Mean dependent var	4.166300
Adjusted R-squared	0.128592	S.D. dependent var	10.13910
S.E. of regression	9.464775	Akaike info criterion	7.343502
Sum squared resid	67365.63	Schwarz criterion	7.392273
Log likelihood	-2782.531	Hannan-Quinn criter.	7.362283
F-statistic	17.00066	Durbin-Watson stat	2.044704
Prob(F-statistic)	0.000000		

Рисунок 3.7 – АРУГ (1,2,3,5,6,11,17)

На рисунку 3.8 зображено прогноз для моделі АРУГ.

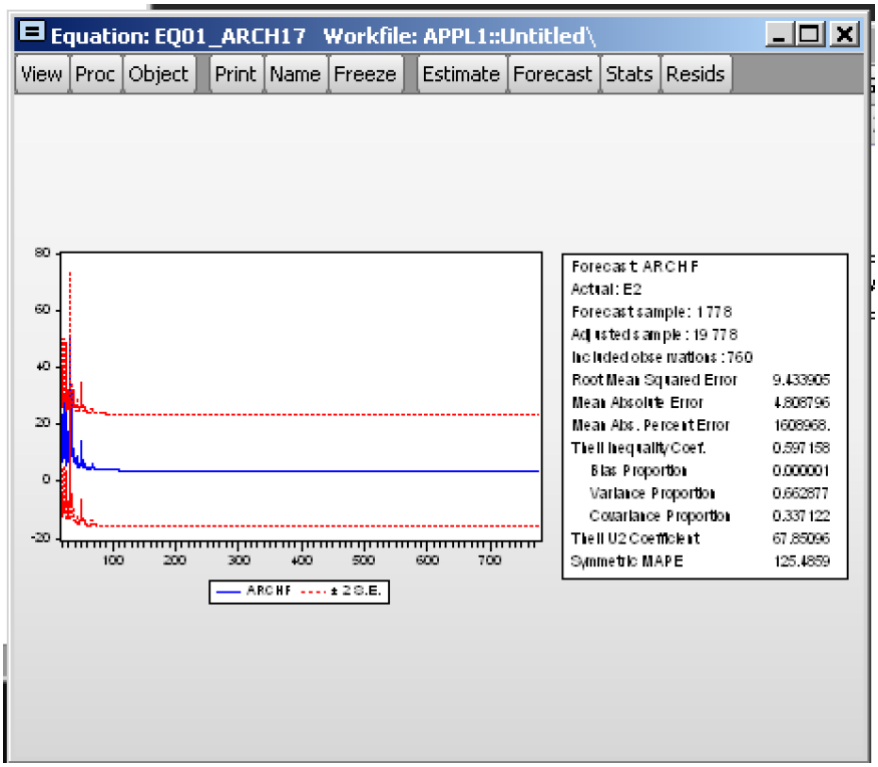

Рисунок 3.8 – значення прогнозу

Як бачимо ми отримали значення R^2 – 0.13, що вказує на те, що ми отримали дуже неточну модель. Також критерій DW = 2.04, ідеальне значення для якого 2. Критерій Тейла – 67.85, а ідеальне значення 0. MAPE = 1608968. З цих даних бачимо, що модель погано описує наші дані.

Побудуємо модель УАРУГ для покращення прогнозування наших даних.

Для початку подивимось на АКФ та ЧАКФ для обчисленого ряду Н.

На рисунку 3.9 зображено АКФ та ЧАКФ для Н.

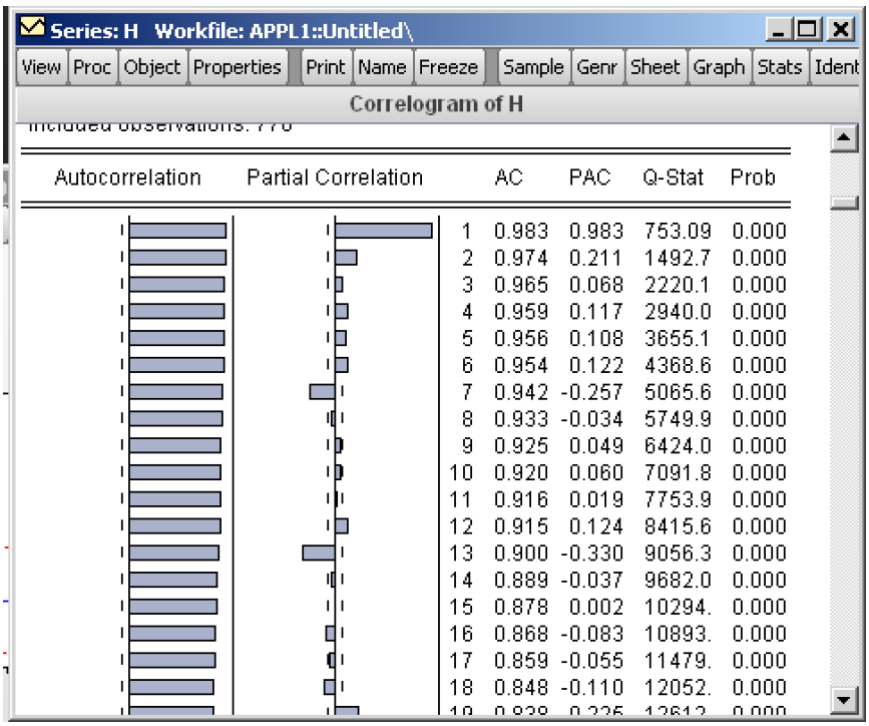

Рисунок 3.9 – АКФ та ЧАКФ для ряду H

Визначивши лаги 1, 2, 7, 12 будуємо модель.

На рисунку 3.10 зображено УАРУГ модель для лагів 1, 2, 7, 12.

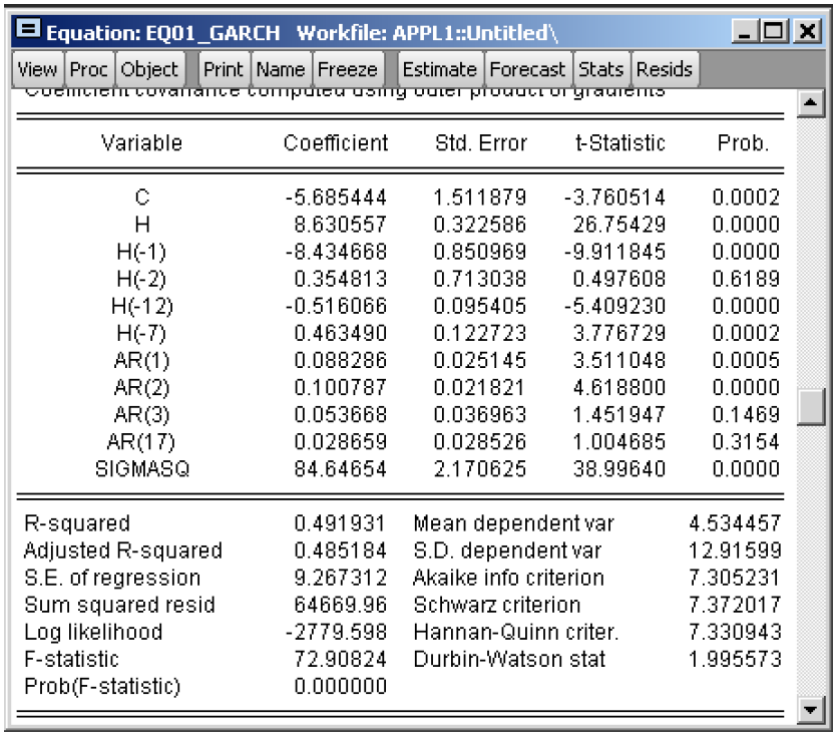

Рисунок 3.10 – УАРУГ (1,2,7,12)

На рисунку 3.11 зображено прогноз для моделі УАРУГ.

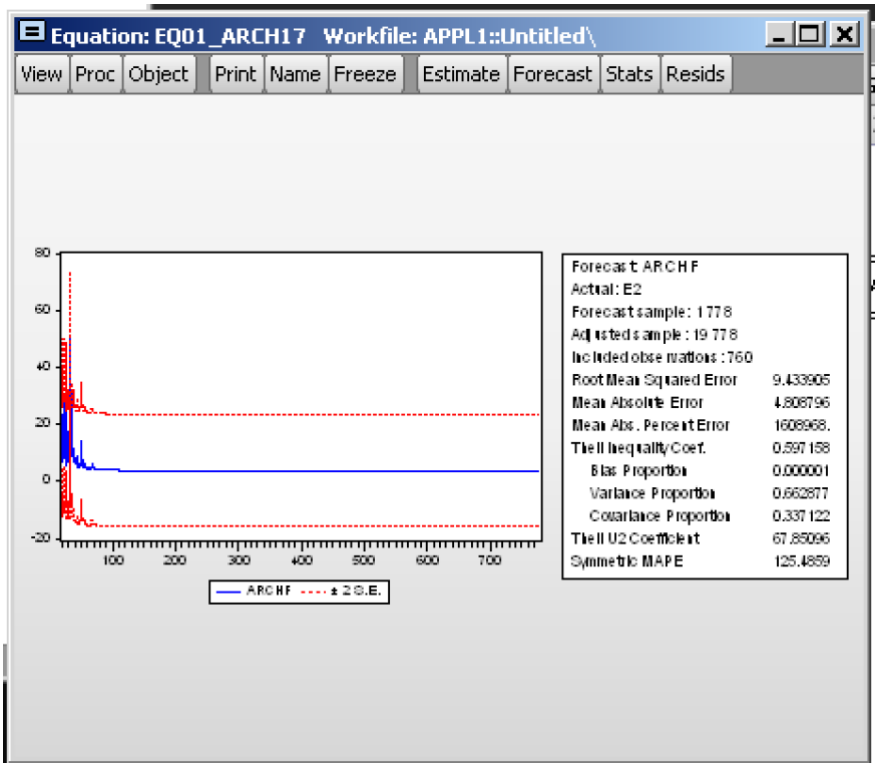

Рисунок 3.11 – значення прогнозу УАРУГ

Як бачимо ми отримали значення R^2 – 0.49, що вказує на те, що модель неточна, але значно краща за АРУГ модель. Також критерій DW = 1.99, що близько до ідеального значення яке дорівнює 2. Критерій Тейла – 50.48, а ідеальне значення 0. MAPE = 1225652. З цих даних бачимо, що модель погано прогнозує наші дані.

Побудуємо модель Е-УАРУГ для покращення прогнозування наших даних.

На рисунку 3.12 зображено Е-УАРУГ модель для лагів 1, 2, 7.

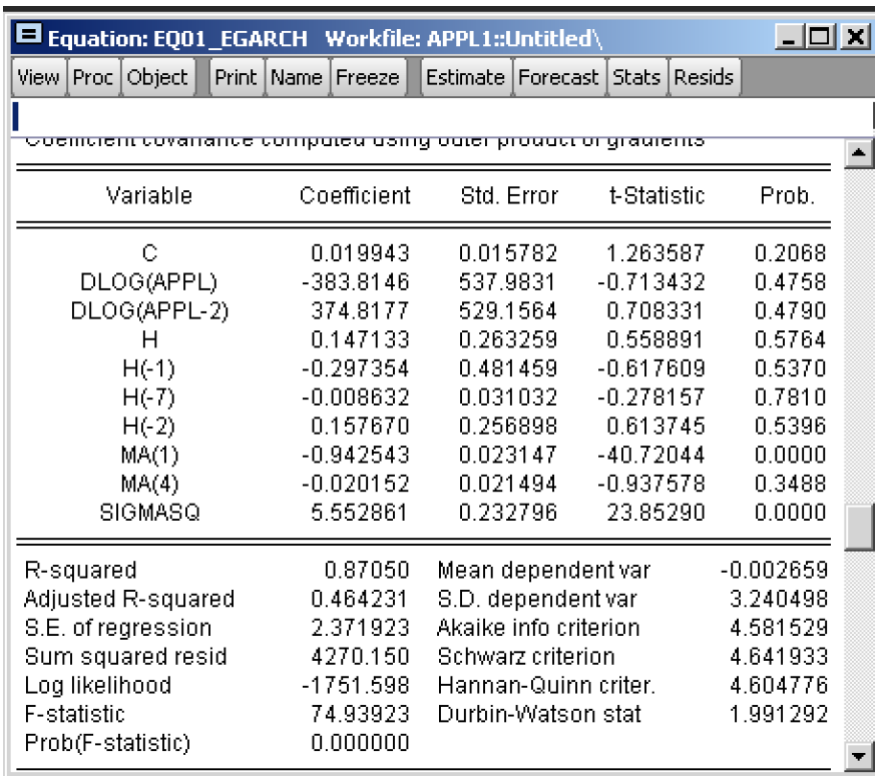

Рисунок 3.12 – УАРУГ (1,2,7,12)

На рисунку 3.13 зображено прогноз для моделі Е-УАРУГ.

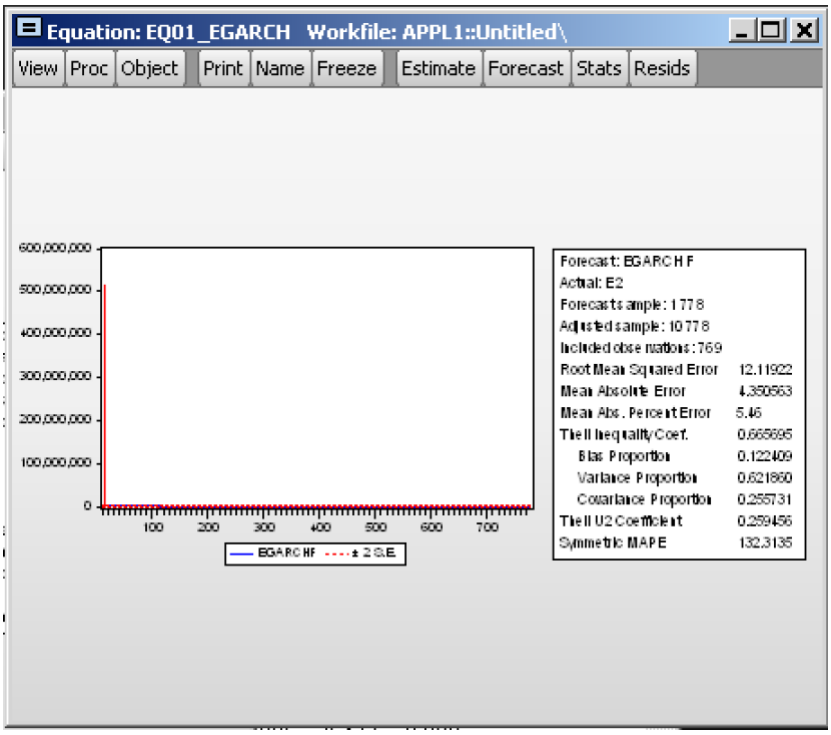

Рисунок 3.13 – значення прогнозу Е-УАРУГ

Як бачимо ми отримали значення R^2 – 0.87, що вказує на те, що модель досить точно описує наші дані і значно краще за УАРУГ, АРУГ моделі. Також критерій DW = 1.99, що близько до ідеального значення яке дорівнює 2. Критерій Тейла – 0.25, а ідеальне значення 0. MAPE = 5.46. З цих даних бачимо, що модель досить гарно прогнозує наші дані.

Проведемо аналогічні дослідження для ціни на золото.

На рисунку 3.14 зображено АРУГ модель для лагів 1, 3, 15, 26, 28.

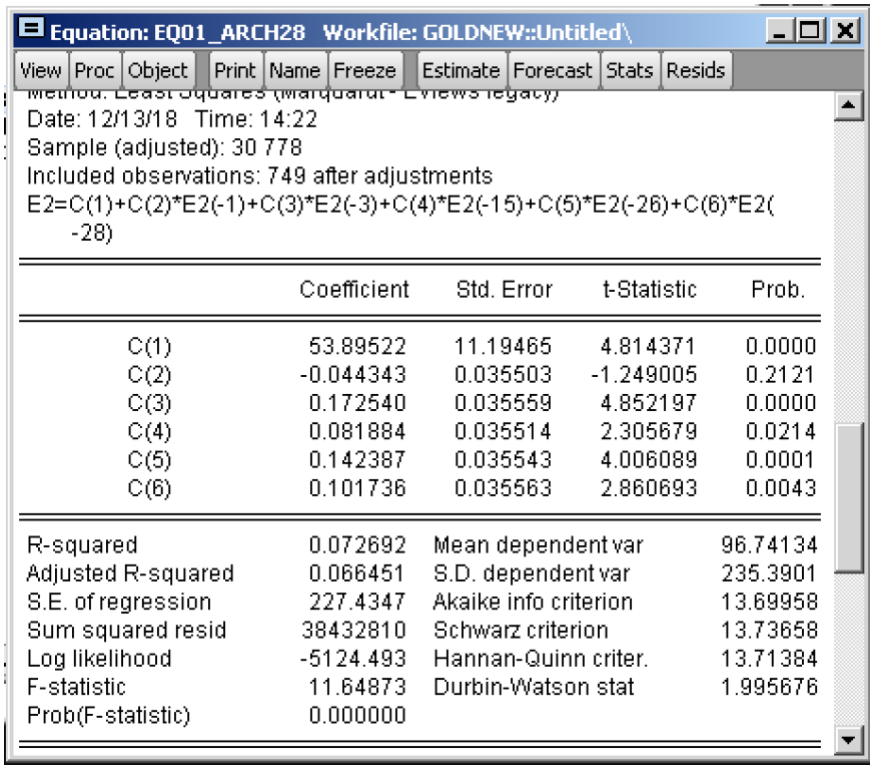

Рисунок 3.14 – АРУГ (1,3,15,26, 28)

На рисунку 3.15 зображено прогноз для моделі АРУГ.

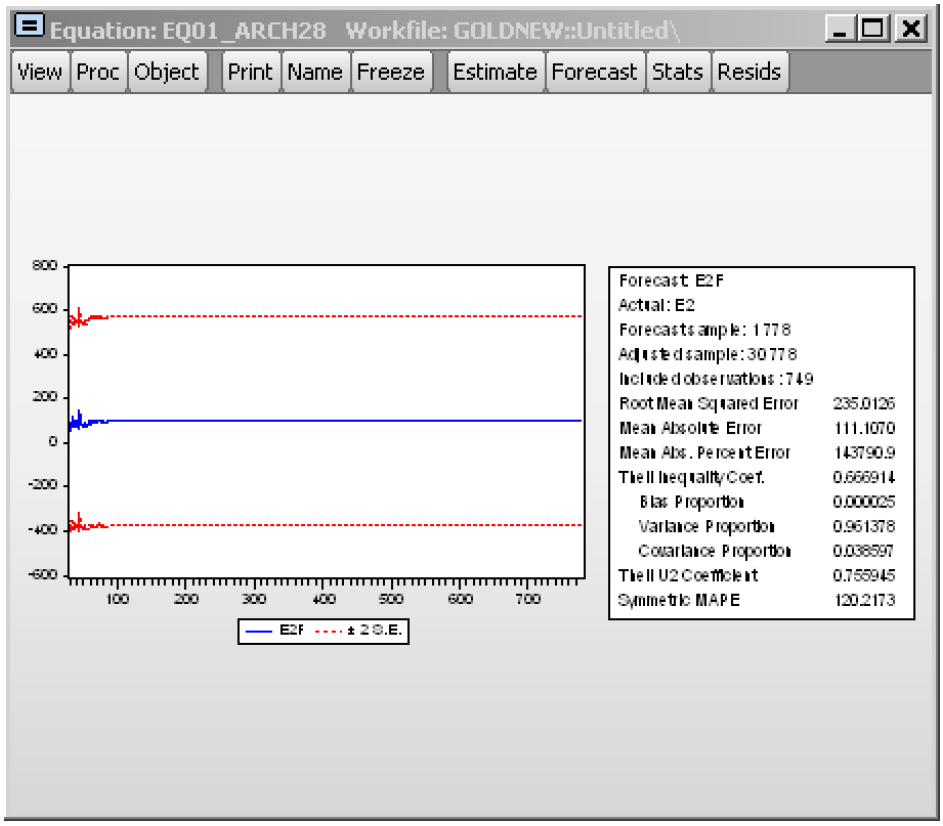

Рисунок 3.15 – значення прогнозу АРУГ

Як бачимо ми отримали значення R^2 – 0.07, що вказує на те, що модель неточна. Також критерій DW = 1.99, що близько до ідеального значення яке дорівнює 2. Критерій Тейла – 0.75, а ідеальне значення 0. MAPE = 143790. З цих даних бачимо, що модель погано прогнозує наші дані.

Побудуємо моделі УАРУГ та Е-УАРУГ для покращення прогнозування наших даних.

На рисунку 3.16 зображено УАРУГ модель для лагів 1, 10, 22, 31.

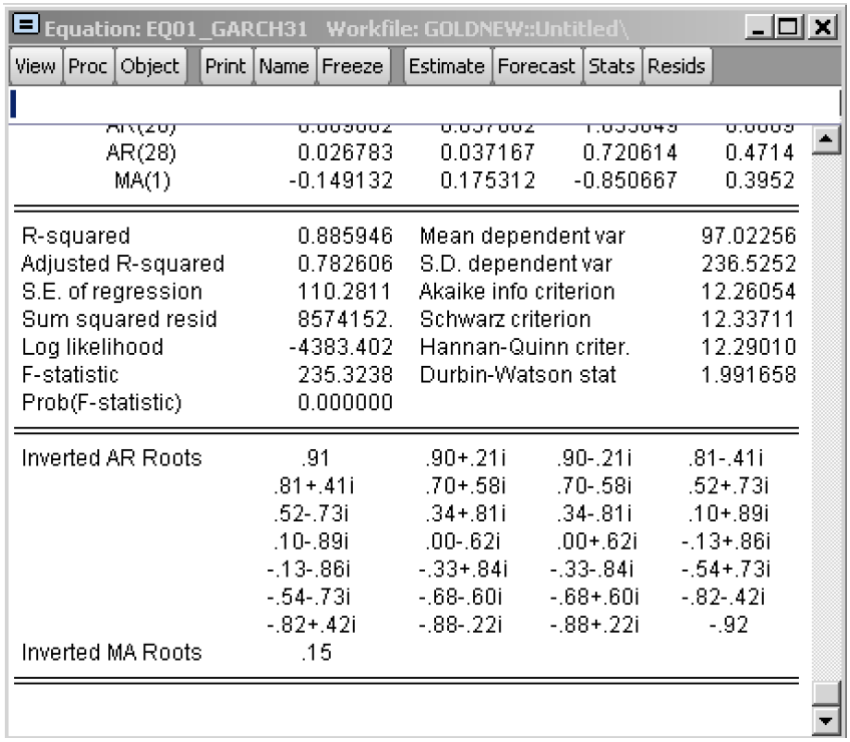

Рисунок 3.16 – УАРУГ (1,10,22,31)

На рисунку 3.17 зображено прогноз для моделі УАРУГ.

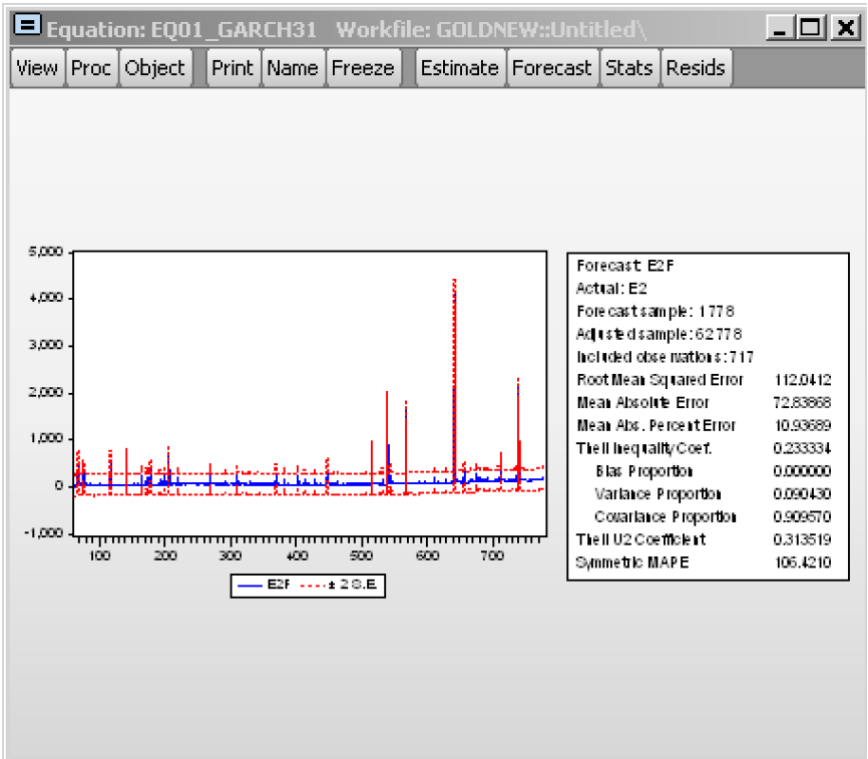

Рисунок 3.17 – значення прогнозу УАРУГ

На рисунку 3.18 зображено Е-УАРУГ модель.

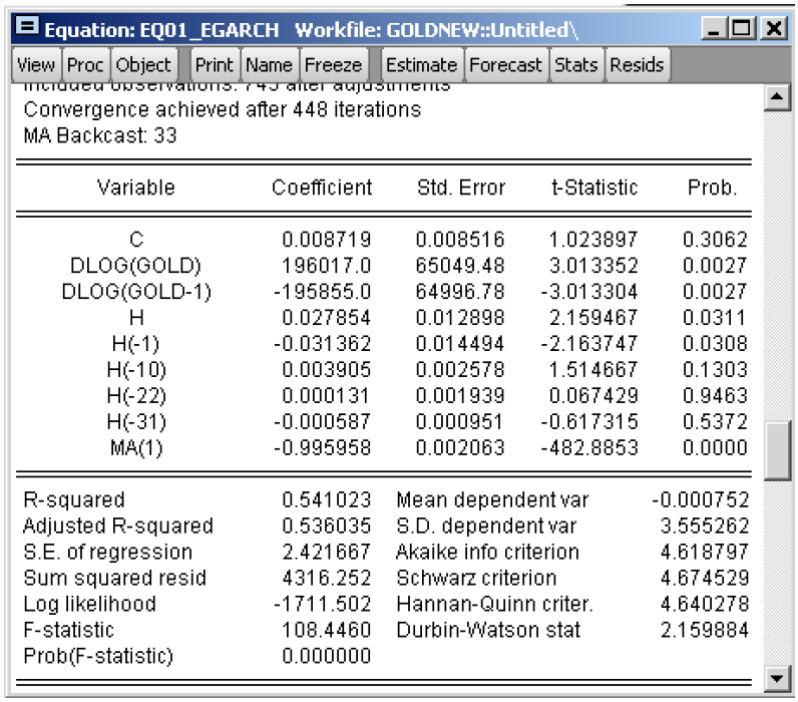

Рисунок 3.18 – Е-УАРУГ

На рисунку 3.19 зображено прогноз для моделі Е-УАРУГ.

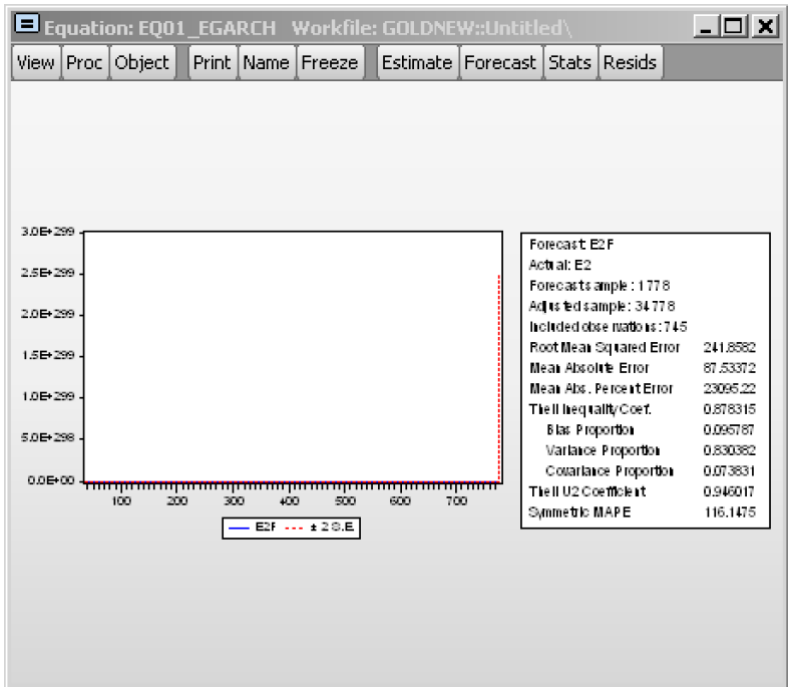

Рисунок 3.19 – значення прогнозу Е-УАРУГ

Як бачимо для найкращої моделі УАРУГ ми отримали значення R^2 – 0.88, що вказує на те, що модель досить точно описує наші дані і значно краще за Е-

УАРУГ, АРУГ моделі. Також критерій DW = 1.99, що близько до ідеального значення яке дорівнює 2. Критерій Тейла – 0.31, а ідеальне значення 0. MAPE = 10.93. З цих даних бачимо, що модель досить гарно прогнозує наші дані.

3.3.3 Порівняльна таблиця

Таблиця 3.1 – Аналіз результатів (курс акацій компанії Apple на біржі NASDAQ за 3 роки)

Ряд	Тип моделі	Адекватність моделі		Характеристика прогнозу	
		R^2	DW	MAPE	U
Акції APPLE	АРУГ(1,2,3,5,6,11,17)	0.13	2.04	1608968	67.85
	УАРУГ (1,2,7,12)	0.49	1.99	1225652	50.48
	Е-УАРУГ (1,2,7,12)	0.87	1.99	5.46	0.25

За результатами, що отримали найкращою є Е-УАРУГ (1,2,7,12).

Таблиця 3.2 – Аналіз результатів (курс акацій компанії Google на біржі NASDAQ за 3 роки)

Ряд	Тип моделі	Адекватність моделі		Характеристика прогнозу	
		R^2	DW	MAPE	U
Акції GOOGLE	АРУГ(1,2,3,8,10,15)	0.11	2.05	1678668	66.35
	УАРУГ (1,3,4,8)	0.51	1.99	1125672	45.21
	Е-УАРУГ (1,2,3,10)	0.91	1.99	5.22	0.23

За результатами, що отримали найкращою є Е-УАРУГ (1,2,7,12).

Таблиця 3.3 – Аналіз результатів (курс акацій компанії Amazon на біржі NASDAQ за 3 роки)

Ряд	Тип моделі	Адекватність моделі		Характеристика прогнозу	
		R^2	DW	MAPE	U
Акції AMAZON	АРУГ(1,2,3,4,7,9)	0.06	2.15	2375698	70.89
	УАРУГ (1,4,5,7)	0.35	1.97	1799651	58.91
	Е-УАРУГ (1,2,6,11)	0.82	1.99	7.22	0.28

За результатами, що отримали найкращою є Е-УАРУГ (1,2,6,11).

Таблиця 3.4 – Аналіз результатів (ціна на золото за 3 роки)

Ряд	Тип моделі	Адекватність моделі		Характеристика прогнозу	
		R^2	DW	MAPE	U
Ціна на золото	АРУГ (1,3,15,26,28)	0.07	1.99	143790	0.75
	УАРУГ (1,10,22,31)	0.88	1.99	10.93	0.31
	Е-УАРУГ (1,10,22,31)	0.54	2.15	23095	0.94

За результатами, що отримали найкращою є УАРУГ (1,10,22,31).

Висновки до розділу

Створена СППР в рамках даної дипломної роботи призначена для моделювання та прогнозування економічних та технічних процесів довільної природи на основі емпіричної вибірки даних.

Запропонована програма задовольняє основним характеристикам СППР: використовує дані і моделі, призначена для надання допомоги ОПР, мета створеної системи – підвищення якості та ефективності рішень.

Відмінною особливістю СППР є те, що аналізу піддаються дані, що містяться в операційній системі. Перевагою є компактність завдяки використанню однієї платформи і оперативність у зв'язку з відсутністю необхідності перезавантажувати дані в спеціалізовану систему. З недоліків можна відзначити звуження кола питань, що вирішуються за допомогою системи.

РОЗДІЛ 4 РОЗРОБКА СТАРТАПУ

4.1 Опис ідеї проекту

Опис ідеї стартам проекту описано в таблиці 4.1.

Таблиця 4.1 - Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Розробка моделі для прогнозування нелінійних нестаціонарних процесів з різними вхідними параметрами	Аудит підприємств	Зменшення корумпованості підприємства
	Організаційне управління	Правильна організаційна структура

Сильні, слабкі та нейтральні характеристики ідеї проекту зображено в таблиці 4.2.

Таблиця 4.2 - Визначення характеристик ідеї проекту

№ п/п	Техніко-економічні характеристики ідеї	Потенційні товари/ концепції конкурентів			W (слабка сторона)	N (нейтр. сторона)	S (сильна сторона)
		SAS	SPSS	Statistica			
1.	Відсутність прив’язки до формату даних	- +		-	+		
2.	Моделювання і прогнозування	-		-			+
3.	Ймов. аналіз	+		+			+
4.	АВС-аналіз	+		+		+	
5.	Відомість бренду	-	+	-	+		

4.2 Технологічний аудит ідеї проекту

Технологічний аудит ідеї проекту наведений у таблиці 4.3.

Таблиця 4.3 - Технологічна здійсненність ідеї проекту

№ п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій
1	Простий у реалізації підхід	Eviews	Технологія наявна і не потребує змін. Потрібно реалізувати алгоритм.	Технологія загальнодоступна
2	Ймовірностний аналіз	Eviews	Необхідно реалізувати алгоритм	Технологія загальнодоступна
3	Моделювання та аналіз результатів	Eviews	Необхідно реалізувати розроблені моделі	Технологія загальнодоступна
Обрана технологія реалізації ідеї проекту: для реалізації проекту обрана мова програмування Eviews.				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

Характеристика потенційного ринку стартап-проекту наведена у таблиці 4.4.

Таблиця 4.4 - Попередня характеристика потенційного ринку стартап-проекту

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	3
2	Загальний обсяг продаж, грн/ум.од	2500 ум.од
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Немає
5	Специфічні вимоги до стандартизації та сертифікації	Немає
6	Середня норма рентабельності в галузі (або по ринку), %	30 %

Характеристика потенційних клієнтів стартап-проекту наведена в таблиці 4.5.

Таблиця 4.5 - Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Аудит організаційної структури підприємств	Середній та великий бізнес, що застосовують системи керування ресурсів підприємства, державні підприємтва, аудиторські компанії	ERP система підприємства, розміри оброблюваних даних, технічні обмеження, бюрократичні обмеження	Ефективність прогнозування Швидка обробка даних Оптимальне використання ресурсів

Можливі загрози для стартап-проекту наведені у таблиці 4.6.

Таблиця 4.6 - Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Невпорядкованість і неповнота даних	Клієнтські бази можуть містити невпорядковані дані і також певні дані можуть бути відсутніми	Додавання модуля попередньої обробки даних
2	Нестача технічних ресурсів	Клієнти можуть мати обмежені локальні технічні ресурси, недостатні для повноцінної роботи системи	Винесення модуля обчислення на сервери компаній-партнерів

Фактори можливостей наведені у таблиці 4.7.

Таблиця 4.7 - Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Хмарні обчислення	Можливість виконання усіх обчислень на віддалених серверах	Пристосування модулів обчислення для роботи на сервері
2	Коригування прогнозу	Можливість коригування прогнозу в режимі реального часу на основі власної бази даних та спорідненої інформації з інтернету	Розробка модулів інтеграції з обліковими системами підприємств.

Проведений ступеневий аналіз конкуренції на ринку зображено у таблиці 4.8.

Таблиця 4.8 - Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Вказати тип конкуренції - олігополія	Присутня невелика кількість фірм. Більшість ринку контролюють фірми-гіганти	Впровадження технологічних інновацій. Кооперація з дослідницькими центрами. Розширення функціоналу та задоволення потреб клієнтів.
2. За рівнем конкурентної боротьби - глобальний	Продукція не залежить від країни чи локалізації клієнта	
3. За галузевою ознакою внутрішньогалузева	Продукт спрямований на роздрібну торгівлю	
4. Конкуренція за видами товарів: - за бажанням	Полягає у випередженні задоволення бажань клієнта	
5. За характером конкурентних переваг - неціновa	Переваги передбачають собою ефективність та різноманіття функціоналу	
6. За інтенсивністю - не марочна	Торгова марка майже немає впливу	

Проведений аналіз конкуренції в галузі зображено у таблиці 4.9.

Таблиця 4.9 - Аналіз конкуренції в галузі за М. Портером

Складові аналізу	Прямі конкуренти в галузі	Потенційні конкуренти
	SAS, SPSS, Statistica	SPSS
Висновки	Контролюють велику частину ринку, мають узагальнені рішення	Спрямовані на малий бізнес, не мають локалізацій для більшості країн Європи

Фактори конкурентоспроможності та їх обґрунтування наведені в таблиці 4.10.

Таблиця 4.10 - Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Інновації	Інноваційні рішення мають забезпечити перевагу нашим клієнтам над конкурентами
2	Функціонал	Функціонал повинен покривати вирішення необхідних задач клієнтів
3	Цінова політика	Вартість продукту відіграє велику роль при виборі системи клієнтом
4	Ресурсоємність	Великі затрати технічних ресурсів можуть спровокувати необхідність залучення додаткових коштів

Порівняльний аналіз сильних та слабких сторін проекту відображено у таблиці 4.11.

Таблиця 4.11 - Порівняльний аналіз сильних та слабких сторін RFS

№ п/п	Фактор конкурентоспро-можності	ББали 1-20	Рейтинг товарів-конкурентів у порівнянні з RFS						
			–3	–2	–1	0	+1	+2	+3
1	Інновації	18				+			
2	Функціонал	12	+						
3	Цінова політика	16			+				
4	Ресурсоємність	3						+	

SWOT-аналіз проекту наведено в таблиці 4.12

Таблиця 4.12 - SWOT-аналіз стартап-проекту

Сильні сторони: розумна цінова політика, функціонал забезпечує рішення більшості задач клієнта	Слабкі сторони: відсутність співпраці з інноваційними центрами,
Можливості: впровадження інноваційних рішень, оптимізація роботи продукту	Загрози: неточність результатів

Альтернативи ринкового впровадження проекту розглянуто в таблиці 4.13.

Таблиця 4.13 - Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Спеціалізовані рішення	Висока	1-3 місяців
2	Хмарний сервіс	Висока	3-6 місяців
3	Узагальнення рішення, вихід на нові сфери ринку	Середня	6-12 місяці

4.4 Розроблення ринкової стратегії проекту

Опис та вибір цільових груп потенційних клієнтів зображено в таблиці 4.14.

Таблиця 4.14 - Вибір цільових груп потенційних споживачів

№ п/п	Опис цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Малий бізнес	Абсолютна готовність в розгляді подібних рішень.	Високий попит	Середня	Вхід в сегмент складний
22	Середній бізнес	Середня готовність. В залежності від виду бізнесу, готовність різниться.	Середній попит	Вище середньої	Вхід в сегмент достатньо складний

33	Великий бізнес	Низький рівень, оскільки у великому бізнесі важче переналаштувати свій організаційний устрій.	Низький	Середня	Вхід в сегмент складний
Які цільові групи обрано: 1,2					

В таблиці 4.15 зображено вибір базової стратегії розвитку.

Таблиця 4.15 - Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
	Розробка та створення додаткових функціональних модулів	Таргетні пропозиції бізнесу, проведення презентації функціональних рішень на ярмарках та конференціях	Відсутність аналогічних до новостворених функціональних модулів у конкурентів	Розробка та удосконалення існуючих модулів на основі потреб ринку та інформації від клієнтів

В таблиці 4.16 наведено визначення базової стратегії конкурентної поведінки.

Таблиця 4.16 - Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки

1	Так	Можливі обидва варіанти	Стандартні функціональні модулі будуть виконувати схожі функції.	Унікальна цінова політика, функціональні інновації, сучасні технології

В таблиці 4.17 наведено визначення стратегії позиціонування.

Таблиця 4.17 - Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
1	Висока якість прогнозування в клієнтській сфері застосування	Розробка та удосконалення існуючих модулів на основі потреб ринку та інформації від клієнтів	Спеціалізовані рішення, хмарні сервіси	Прогнозування, передбачення, аналіз

4.5 Розроблення маркетингової програми стартап-проекту

В таблиці 4.18 представлені ключові переваги концепції потенційного товару.

Таблиця 4.18 - Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Широкий функціонал	Вирішення задач	Забезпечує вирішення більшої кількості задач бізнесу
2	Спеціалізовані рішення	Вирішення задач	Забезпечує більш ефективне вирішення задач у звуженій сфері застосування
3	Технічні ресурси	Хмарні сервіси	Дозволяє користуватись рішенням за рахунок віддалених технічних потужностей

Опис трьох рівнів моделі товару відображено у таблиці 4.19.

Таблиця 4.19 - Опис трьох рівнів моделі товару

<table>
<tr><td>Рівні товару</td><td colspan="3">Сутність та складові</td></tr>
<tr><td>I. Товар за задумом</td><td colspan="3">Обробка, аналіз даних. Прогнозування та передбачення потреб споживача</td></tr>
<tr><td rowspan="5">II. Товар у реальному виконанні</td><td>Властивості/характеристики</td><td>М/Нм</td><td>Вр/Тх /Тл/Е/Ор</td></tr>
<tr><td>Швидкодія
Ефективність
Користувацький інтерфейс</td><td>Нм
Нм
Нм</td><td>Тх/Тл/Е
Тх/Тл
Е</td></tr>
<tr><td colspan="3">Якість: стандарти відповідні до законодавства. Створені функціональні скріпти.</td></tr>
<tr><td colspan="3">Пакування: Власний сайт</td></tr>
<tr><td colspan="3">Марка: CM Solutions</td></tr>
<tr><td rowspan="2">III. Товар із підкріпленням</td><td colspan="3">До продажу: застосунок для інтеграції в існуючі системи керування підприємством для прогнозування та передбачення потреб споживачів на основі великих масивів даних</td></tr>
<tr><td colspan="3">Після продажу: Швидкодія, ефективність, легкість у користуванні</td></tr>
<tr><td colspan="4">Закритий код. Захищений від можливості декомпіляції.</td></tr>
</table>

Визначення меж встановлення ціни показано в таблиці 4.20.

Таблиця 4.20 - Визначення меж встановлення ціни

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
1	-	200$/міс	Рівень доходів підприємств надзвичайно високий	150-200$/міс

Формування системи збуту зображено в таблиці 4.21.

Таблиця 4.21 - Формування системи збуту

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Таргетні пропозиції для компаній	Презентації функціоналу	-	-

Концепція маркетингових комунікацій відображена у таблиці 4.22.

Таблиця 4.22 - Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
1	Середній бізнес – оптимальні рішення за невисоку ціну	Соціальні мережі, внутрішньо ринкова комунікація	Прогнозування покупок споживача	Короткий опис переваг продукту, заохочення дізнатись більше	Передбачення покупок споживачів

2	Великий бізнес – повноцінні рішення для покращення продажів	Таргетні дзвінки до клієнтів	Прогнозування покупок споживача	Донести інформацію про оптимальність рішення для бізнесу клієнта	Передбачення покупок споживачів

Висновки до розділу

Отже, відповідно до вищенаведених результатів, можна стверджувати про наявність попиту на запропоновану систему. Варто зауважити, що присутня мала конкуренція, оскільки рішення нове, тож інноваційна складова продукту дозволяє суттєво збільшити конкурентоспроможність проекту.

ПЕРЕЛІК ПОСИЛАНЬ

1. Химмельблау Д. Анализ процессов статистическими методами / Химмельблау Д. – М.: Мир, 1973. – 957 с
2. Згуровский М.З. Аналитические методы калмановской фильтрации для систем с априорной неопределенностью / Згуровский М.З., Подладчиков В.Н. – К.: Наукова думка, 1995. – 298 с.
3. A Tutorial on Particle Filters for on-line Non-linear Non-Gaussian Bayesian Tracking / [Arulampalam, S., Maskell, S., Gordon, N., Clapp, T.] // IEEE Trans. Signal Processing. – 2001. – vol. 50. – pp. 174-188.
4. Gordon N. J. Novel approach to nonlinear/non-Gaussian Bayesian state estimation / Gordon N. J., Salmond D. J., Smith A. F. M. // IEE Proceedings-F. – 1993. – vol. 140, No. 2. – pp. 107-113.
5. Brown R.G. Smoothing forecasting and prediction if discrete time series / Brown R.G. – New York: Courier Corporation, 1963. – 468 p.
6. Бідюк П.І. Проектування комп’ютерних інформаційних систем підтримки прийняття рішень: Навчальний посібник. / Бідюк П.І., Коршевнюк Л.О. – К.: ННК «ІПСА» НТУУ «КПІ», 2010. – 340 с.
7. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов / Лукашин Ю.П. – М.: Финансы и статистика, 2003. – 413 с.
8. Бідюк П. І. Часові ряди: моделювання і прогнозування / Бідюк П. І., Савенков О. І., Баклан І.В. – К.: ЕКМО, 2003. – 144 с.
9. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking / [M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp]. // IEEE. Transactions on Signal Processing. – 2002. – Vol. 50, No. 2. - pp. 174 – 188.
10. F. Gustafsson. Particle filter theory and practice with positioning applications / F. Gustafsson // IEEE. Aerospace and Electronic Systems Magazine. – 2010. - Vol. 25, No. 7. – pp. 53–82.

11. Бідюк П.І. Меднтоди прогнозування / Бідюк П.І., Менайленко О.С., Половцев О.В. – Луганськ: Альма Матер, 2008. – 605 с
12. Дрейпер Н. Прикладной регрессионный анализ / Дрейпер Н., Смит Г. – М.: Финансы и статистика, 1986. – 366 с.
13. Герасимов Б. М. Человеко-машинные системы принятия решений с элементами искусственного интеллекта / Герасимов Б. М. , Тарасов В. А., Токарев И. Б. – К.: Наукова Думка, 1993. – 184 с.
14. Бидюк П.И. Временные ряды: моделирование и прогнозирование / П.И. Бидюк, О.И. Савенков, И.В. Баклан. — К.: ЕКМО, 2003. — 141с.
15. Бокс Дж., Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. - 402 с.
16. Бідюк П.І. Принципи прогнозування часових рядів / Бідюк П.І., Шехтер Д.В., Клименко О.М. // Наукові вісті НТУУ „КПІ”. – 2005. – № 5. – С. 14-25
17. Бидюк П.И. Курс лекций по анализу временных рядов [Текст]. / П.И. Бидюк – К.: НТУУ «КПИ», 2009. – 450 с.
18. Анфилатов В.С. Системный анализ в управлении / Анфилатов В.С., Емельянов А.А, Кукушкин А.А. – М.: Финансы и статистика, 2002. – 368 с.
19. Ларичев О.И. Теория и методы принятия решени / Ларичев О.И. – М.: Логос, 2000. – 296 с.
20. Нейман Дж. Теория игр и экономическое поведение / Пер. с англ. / Нейман Дж., Моргенштерн О. – М.: Наука, 1970. – 707 с.
21. Little J.D.C. Models and Managers: The Concept of a Decision Calculus / Little J.D.C. // Management Science. – 1970. – Vol.16, No 8. – pp. 10-26.
22. Power D.J. A Brief History of Decision Support Systems [Електронний ресурс]. – 2003. – Режим доступу до ресурсу: http://dssresources.com/history/dsshistory.html
23. Небава М.І. Теорія макроекономіки: Навч. посіб. / М.І. Небава. – К.: Слово, 2005. – 536 с.

24. Згуровский М. З. Основы вычислительного интеллекта : монография / М. З. Згуровский, Ю. П. Зайченко. – К.: Наукова думка, 2013. - 406с.

25. Зайченко Ю.П. Исследование операций: Учеб. пособие / Зайченко Ю.П.– К.: Слово, 2003. – 688 с.

26. Калина А. В. Современный экономический анализ и прогнозирование (макро- и микроуровень): Учебно-методическое пособие. / Калина А. В., Конева М. И., Ященко В. А. – К.: МАУП, 1997. -272 с.

27. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования/ Ю.П. Лукашин – М.: Финансы и статистика, 2003. – 414 с.